Self-Consciousness in Kant's Moral Philosophy

Self-Consciousness in Kant's Moral Philosophy

A dissertation presented

by

James Bondarchuk

to

The Philosophy Department

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Philosophy

Harvard University

Cambridge, Massachusetts

July 2018

Dissertation Advisor: Professor Christine Korsgaard James Bondarchuk

Self-Consciousness in Kant's Moral Philosophy

Abstract

In the *Critique of Practical Reason* (1788), Kant declares that our consciousness of the moral law is a "fact of reason," and that this fact suffices to establish the reality of moral obligation. With this doctrine, Kant asserts that a "deduction" of morality, such as he attempted in the *Groundwork of the Metaphysics of Morals* (1785), is neither necessary nor possible. This reversal has seemed to some commentators to be a retreat to the pre-critical dogmatism that the *Critique of Pure Reason* (1781, 1787) positioned itself against. I defend the doctrine of the fact of reason against this charge, arguing that the doctrine is not just consistent with, but in fact an expression of, the fundamental methodological commitments of critique. The lynchpin of my defense is a conception of the critical method according to which critique relies on a form of self-cognition that has its basis in the subject's pure apperception. In chapter one, I give an account of pure apperception according to which it is the subject's consciousness of engaging in the activity of thought. In chapter two, I argue that the pure apperceptive form of thought makes available a form of self-cognition, labeled "reflection," on the basis of which critique can be undertaken. Crucially, critique's presumption of our rational autonomy includes the presumption of a capacity for reflective self-cognition. In chapter three, I argue that the self-consciousness of autonomous agency is identical to the pure apperception of moral deliberation and action. This fact, along with the moral law's status as a fundamental practical principle, entails that a deduction of the moral law is impossible. Finally, in chapter four, I argue that Kant's invocation of a "fact of reason" indicates the second *Critique*'s reliance on a distinctively practical mode of reflection. This is no lapse into dogmatism, but rather an expression of critique's commitment to the autonomy of reason.

Table of Contents

Acknowledgments

This dissertation would not have been possible without the outstanding support and guidance of the members of my committee: Chris Korsgaard, Matt Boyle, and Béatrice Longuenesse.

Chris Korsgaard was my principal advisor throughout my tenure as a graduate student at Harvard. It was during a meeting with her in my second year that I first floated the core idea that I would expand into this dissertation. (It was originally to make up *the second half* of my second year paper, but the rubber of unrealistic ambition eventually met the road of an all-too-real deadline. One of Chris's chief virtues as an advisor is that she encourages bold, interesting projects.) Chris was a wonderful sounding board for ideas that were still in the development stage. She has a seemingly boundless capacity for philosophical imagination. She also has an uncanny ability to identify the deep philosophical issues that underlie some arcane, technical question: Chris often understood my project better than I did, and knew exactly what to ask to bring those issues into relief. In addition to the specific ways that she influenced this project, I am grateful to Chris for her many years of investment in my intellectual development. The habits of mind I cultivated under her guidance have made me a better thinker and more humane person.

Matt Boyle was the advisor with whom I met most often to discuss this project. Our conversations about Kant's system, which sometimes ran several *hours* in length, helped me avoid many wrong turns, and shaped this project in extremely detailed and subtle ways. Matt always conveyed a note of respect for my ideas. Even when he disagreed with me on some particular point, he would go out of his way to acknowledge those considerations that militated in favor of my view. His generous and open-minded approach to advising is, I believe, a reflection of broader qualities: he cares deeply about philosophy, has an admirable tolerance for uncertainty, and is utterly lacking in

pretension. I am most grateful to Matt for the kindness he showed me in moments when my self-confidence was in short supply.

Béatrice Longuenesse joined my committee after the prospectus stage, but her impact was felt from the beginning. I owe a great deal of my understanding of the *Critique of Pure Reason* to her book *Kant and the Capacity to Judge*, which I first encountered when I was a masters student at the University of Wisconsin-Milwaukee. Many of the ideas contained in this dissertation began to take shape in the Fall 2013 iteration of Harvard's Kant Reading Group, which we devoted to that book. Since joining my committee, Béatrice has been the impetus for substantial improvements to this project. Her extremely thorough feedback helped me to clarify and refine many of my central claims, and brought my discussion of the first *Critique* to a higher level of sophistication. I would also like to thank Béatrice for her personal encouragement as I neared completion.

I am very fortunate to have been able to attend the aforementioned reading group, which was organized by Matt Boyle and Farid Masrour. In addition to Matt and Farid, I would like to thank the following people for our many stimulating discussions about Kant's philosophy: Charles Parsons, Yoon Choi, Rachel Achs, Chandler Hatch, Thomas Pendlebury, and Sandy Diehl.

I benefited tremendously from the feedback I received in the department's Moral and Political Philosophy Workshop. I received particularly helpful comments from Tim Scanlon, Selim Berker, Jeremy Fix, Doug Kremm, Byron Davies, and Olivia Bailey.

Finally, I would like to express my highest gratitude to Pardis Dabashi, who knows what she did.

Abbreviations of Kant's Works

References to Kant's works employ the following abbreviations, followed by the volume and page numbers of the Prussian Academy Edition of Kant's Complete Writings. One exception is the *Critique of Pure Reason*, in which case page numbers are preceded by 'A' or 'B' to indicate the original 1781 or 1787 edition, respectively.

An	*Anthropology from a Pragmatic Point of View*, trans. Robert B. Louden. In *Anthropology, History, and Education: The Cambridge Edition of the Work of Immanuel Kant*, pp. 227-429. New York: Cambridge University Press, 2007.
G	*Groundwork of the Metaphysics of Morals*, trans. Mary Gregor. Cambridge: Cambridge University Press, 1997.
JL	*The Jäsche Logic*. In *Lectures on Logic: The Cambridge Edition of the Works of Immanuel Kant*, trans. J. Michael Young, pp. 517-640. New York: Cambridge University Press, 1992.
KpV	*Critique of Practical Reason*. In *Practical Philosophy: The Cambridge Edition of the Works of Immanuel Kant*, trans. Mary Gregor, pp. 133-258. New York: Cambridge University Press, 1996.
KrV	*Critique of Pure Reason*, trans. Paul Guyer and Allen W. Wood. New York: Cambridge University Press, 1998.
KU	*Critique of the Power of Judgment*, trans. Paul Guyer and Eric Matthews. New York: Cambridge University Press, 2000.
LE	*Lectures on Ethics*, ed. Peter Heath and J.B. Schneewind, trans. Peter Heath. New York: Cambridge University Press, 1997.
MM	*The Metaphysics of Morals.* In *Practical Philosophy: The Cambridge Edition of the Works of Immanuel Kant*, trans. Mary Gregor, pp. 353-603. New York: Cambridge University Press, 1996.
NF	*Notes and Fragments: The Cambridge Edition of the Works of Immanuel Kant*, trans. Bowman, Guyer, and Rauscher. New York: Cambridge University Press, 1997.
Prol	*Prolegomena to Any Future Metaphysics*, trans. Gary Hatfield. New York: Cambridge University Press, 1997.
Rel	*Religion Within the Boundaries of Mere Reason*. In *Religion and Rational Theology: The Cambridge Edition of the Works of Immanuel Kant*, trans. Allen Wood and George di Giovanni, pp. 39-215. New York: Cambridge University Press, 1996.

Dedicated to Tom Fehse.

INTRODUCTION

In the Preface to the *Groundwork of the Metaphysics of Morals* (1785), Kant announces his intention to undertake "the search for and establishment of the *supreme principle of morality*" (G 4:392). If the first two chapters of the *Groundwork* can be said to constitute the "search" for the supreme principle of morality—the identification of the implicit notion of duty operative in everyday moral reasoning (*Groundwork* I), and the philosophical elucidation of that concept (*Groundwork* II)—then the third and final chapter might be called its "establishment." For it is there that Kant argues that all rational agents are bound by the moral law as an unconditionally authoritative practical principle.

The first two chapters of the *Groundwork* take for granted that we are morally obligated. Their aim is to demonstrate what the structure and content of morality must be, assuming there is such a thing, through the analysis of moral concepts. On the basis of such analysis, Kant argues that moral precepts issue from a single, fundamental principle of pure practical reason.[1] The first formulation of this "supreme principle," known among commentators as the "Formula of Universal Law," enjoins rational agents such as ourselves to act only on those principles whose universal adoption we could concomitantly legislate (4:421).[2] Additionally, Kant argues that in choosing principles on the basis of their suitability for universal adoption, we thereby exercise our rational autonomy. Thus, for Kant, we act autonomously whenever we act according to the fundamental normative standard of our own pure practical reason.

---

[1] The concept of pure practical reason and its relationship to a notion of autonomy will be explained in some detail in chapter three.

[2] In the *Groundwork*, this formulation is expressed as follows: "act only in accordance with that maxim through which you can at the same time will that it become a universal law" (4:421). In the *Critique of Practical Reason*, it is expressed somewhat differently: "So act that the maxim of your will could always hold at the same time as a principle in a giving of universal law" (KpV 5:30). Kant offers two additional formulations of the categorical imperative, each of which he regards as a different articulation of a common principle. The second formulation (labeled the 'Formula of Humanity' by commentators) is: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (G 4:429). The third formulation (labeled the 'Formula of the Kingdom of Ends' by commentators) is: "every rational being must act as if he were by his maxims at all times a lawgiving member of a universal kingdom of ends" (4:438).

Of course, Kant understands that the analysis of moral concepts is, by itself, no justification for the claim that we are actually morally obligated. To allay the worry that morality might be a "chimerical idea without any truth" (G 4:445), *Groundwork* III aims to show that the Formula of Universal Law is categorically binding on all finite rational agents. This is the task of a "deduction of the supreme principle of morality" (G 4:463, my emphasis). In brief, Kant argues that every rational agent is, just in virtue of being a rational agent, committed to her own autonomy—committed, that is, to a conception of herself as capable of acting on the basis of pure reason alone. But if that is correct, then she is likewise committed to the Formula of Universal Law as her categorical imperative, because the Formula of Universal Law is the principle of autonomy.[3]

Whatever the individual merits of his attempted "deduction" of morality, Kant would later significantly change his mind about the place of any such argument within his greater system. In the *Critique of Practical Reason* (1788), Kant claims that "the objective reality of the moral law cannot be proved by any deduction" (KpV 5:47). Given that Kant had earlier felt it necessary to *justify* our self-conception as moral agents, one might have expected him to conclude that we are not warranted in regarding ourselves as morally obligated. But this is not what he does. Instead, Kant asserts that our consciousness of the imperatival force of morality is a "fact of reason," and that such consciousness suffices to establish the reality of moral obligation (5:31; 5:42ff.). On this view, moral consciousness entitles us to regard ourselves as morally obligated in lieu of even the *possibility* of a proof of such obligation. Thus, Kant's later view is not just that a deduction of the moral law is impossible, but that it is unnecessary.

Unsurprisingly, Kant's apparent stipulation of a "fact" in place of a deduction has won few supporters. Even among sympathetic readers, the doctrine of the fact of reason has seemed like an

[3] Here I ignore the complication that morality is felt as an *imperative* only by agents that are subject to non-rational desires and impulses. Cf. G 4:414.

instance of precisely the sort of "dogmatic" metaphysics that the *Critique of Pure Reason* (1781, 1787) inveighed against. With the first *Critique*, Kant aimed to show the source and limits of synthetic, *a priori* cognition. But that project took place against a background of uncertainty regarding the ability of any finite rational subject to know synthetic, *a priori* truths. By Kant's own lights, what distinguished critique from the dogmatism of his rationalist predecessors was precisely the attempt to *validate*, as well as determine the limits of, our capacity for synthetic, *a priori* cognition. The broad tenets of that attempt are well known: Kant would conclude that certain *a priori* principles are both necessary for cognition and indeed applicable to objects that we perceive in space and time. Nevertheless, space and time are the merely ideal forms of our own sensible intuition, and we lack any capacity for cognition of things-in-themselves. Against this background, it would appear that the *Groundwork*'s apparent need to prove the synthetic, *a priori* claim that we are morally obligated is an expression of the same critical demand. For this reason, Kant's subsequent announcement of an unargued-for "fact of reason" seems quite antithetical to the method of critique. Furthermore, it fuels the suspicion that Kant declared that a deduction of the moral law is impossible only because of his own failed attempt to provide one.

The goal of this dissertation is to defend Kant's doctrine of the fact of reason against this line of attack. To that end, I will provide defenses of both the claim that a deduction of the moral law is impossible as well as the claim that moral consciousness entitles us to regard ourselves as genuinely morally obligated, i.e., that a deduction of the moral law is unnecessary. Regarding the latter, I will argue that Kant's invocation of the fact of reason in place of a deduction is rooted in the fundamental methodological commitments of critique. But I will arrive at these conclusions by way of an unusual route: the theory of self-consciousness that Kant articulates in his *Critique of Pure Reason*. For this reason, the first half of this dissertation will concern Kant's theoretical philosophy, specifically Kant's doctrine of pure apperception and its connection to the concept and method of a critique of theoretical reason. These chapters will serve to illustrate broad parallels with Kant's practical

philosophy that bear, in ways that are not readily apparent, on whether a deduction of the moral law can or must be given.

The outline for this dissertation is as follows.

In chapter one, I explicate Kant's notion of *pure apperception*. According to my interpretation, pure apperception is the subject's consciousness of herself as the "agent" of the activity of thinking, and this self-consciousness is a constitutive feature of thought. I argue that Kant invokes an agential conception of self-consciousness in order to explain how consciousness of oneself as a thinking subject is possible despite his insistence that to be conscious of oneself in this way is not to relate to oneself as an object of representation. What distinguishes "pure" from what Kant calls "empirical" apperception is just this: pure apperception is the subject's consciousness of *combining representations according to a certain rule, or "function," of unification*. This turns out to be identical to the subject's consciousness of engaging in the activity of thought, as opposed to the consciousness of some given empirical content. In this way, Kant's doctrine of pure apperception can be regarded as a response to broadly Humean doubts about the possibility of self-consciousness.

In chapter two, I relate the above account of pure apperception to the method of critique as it is exhibited in the *Critique of Pure Reason*. The purpose of a critique of pure reason is to determine the source and legitimacy of certain principles of pure reason according to reason's own standards of critical self-assessment. Importantly, however, critique cannot proceed except on the basis of certain assumptions about the nature of our capacity for thought, including our capacity for the discursive cognition of objects. Kant believes we can achieve genuine self-cognition and therefore be justified in these assumptions by engaging in *reflection* on those capacities. The activity of reflection is a mode of the capacity for pure apperception described in chapter one. Whereas pure apperception is usually a spontaneous and implicit feature of ordinary thought, reflection is the activity whereby the subject deliberately attends to her thinking activity in order to achieve *explicit* self-cognition. Among the items

of self-cognition grounded in reflection is the doctrine of pure apperception itself. With respect to the overarching aims of the dissertation, the most significant conclusion of this chapter is that Kant's reliance on reflection has its basis in his commitment to a conception of critique as exhibiting the autonomy of reason. I argue that an autonomous rational capacity must be self-conscious in the pure apperceptive sense. From this it follows that the assumption of our rational autonomy includes the assumption of a capacity for reflective self-cognition. As a consequence, reflection is not a mere epistemic enabling condition for critique's operation but rather tied to critique in conception.

In chapter three, I take up Kant's claim that a deduction of the moral law is impossible. This chapter refers back to the doctrine of pure apperception presented in chapter one. I argue, first, that autonomous agency must be an essentially self-conscious capacity: in exercising that capacity, we must be at least implicitly conscious of ourselves *as autonomous agents*. Next, I argue that Kant holds that the self-consciousness of autonomous agency must be a form of pure apperception. This is because, as with the consciousness of ourselves as thinking subjects, the relevant form of self-consciousness cannot be understood on the model of our cognition of objects. Thus, in trying to explain the possibility of the self-consciousness of autonomous agency, we face an explanatory constraint that is formally identical to the one we encountered in chapter one, and we should therefore expect the form of Kant's solutions to be the same. In particular, the self-consciousness of autonomous agency consists in the pure apperception of specifically *moral* deliberation and action, that is, of choosing principles on the basis of their suitability for universal adoption, as determined by the Formula of Universal Law. On this basis, I claim that the self-consciousness of autonomous agency turns out to be identical to Kant's "fact of reason," i.e., the consciousness of the imperatival force of morality in practical deliberation. But if that is right, then we cannot appeal to the agent's consciousness of her autonomy in practical deliberation as the basis for a deduction of morality. I conclude this chapter by showing that the moral law's status as a *fundamental practical principle* ultimately precludes the possibility

of any other path to the claim that the moral law is the principle of our will. Therefore, a deduction of the moral law is impossible.

Finally, in chapter four, I take up the question of whether a deduction of the moral law is necessary—whether, that is, our consciousness of moral requirements in practical deliberation entitles us to regard ourselves as moral agents. This chapter refers back to the account of critique offered in chapter two, where I argued that critique necessarily relies on items of self-cognition made available through reflection on our discursive powers. Since the fact of reason is our pure apperception of moral deliberation and action, Kant's invocation of this "fact" as the basis for a critique of practical reason indicates his reliance on a distinctively *practical* mode of reflection. That is, just as the critique of pure theoretical reason proceeded on the basis of reflection on our capacity for the discursive cognition of objects, so the critique of practical reason is made possible by reflection on our capacity for pure practical deliberation and action. If that is right, then Kant's reliance on the "fact" of moral consciousness is, taken by itself, consistent with a suitably generalized conception of the critical method. But this poses the following question: Why wasn't Kant similarly entitled to declare our consciousness of deploying various synthetic, *a priori* principles of *theoretical* cognition a "fact of reason"? Why, for example, couldn't Kant make such consciousness the basis for an assertion that, necessarily, every event has a cause? In brief, Kant's answer is that in the case of pure practical reason, the question of the *application* of synthetic, *a priori* principles to objects given in sensibility does not arise. In other words, establishing the objective validity of the moral law does not require anything that is analogous to the first *Critique*'s task of proving that sensibly-given objects are constituted in such a way that principles of the understanding necessarily apply to them. For that reason, the pure apperception of moral agency *suffices* to establish the objective validity of morality. This is the sense in which the "fact of reason" supplants the need for a deduction of the moral law.

As with the critique of pure theoretical reason, the second *Critique*'s presumption of a capacity for reflective self-cognition has its basis in the presumption of the normative autonomy of reason. It is a basic commitment of critique that pure reason is ultimately normatively answerable only to itself, and reason exhibits its normative autonomy in its assumption that we are aware, in pure apperception, of the nature of our own rational activity. On this conception of critique, there is no attempt to *validate* that which is cognized through pure apperception—to prove that what we regard through pure apperception as the principle of our activity, whether theoretical or practical, is *in fact* the principle of our activity. But notice that this feature of critique implies that the *Groundwork*'s attempt to provide a deduction of the moral law actually represents a significant *departure* from the critical method. The third chapter of the *Groundwork* is concerned to answer, in effect, a challenge that has its basis in a perspective alienated from that of self-conscious moral deliberation, namely, to prove that rational agents are in fact capable of pure practical reason—that morality is not a "chimerical idea." In this way, *Groundwork* III pursues the deeply *uncritical* project of attempting to validate pure practical reason according to a standard other than that of pure reason itself. In the *Critique of Practical Reason*, by contrast, it is assumed that the fundamental principle of our will is an item of reflective self-cognition. This is no lapse into pre-critical dogmatism, but rather an expression of critique's most fundamental normative commitment.

CHAPTER ONE

Kant's Theory of Self-Consciousness

§1.0 The aim of this chapter is to present and defend an interpretation of the theory of self-consciousness that Kant articulates in his *Critique of Pure Reason*. To that end, I will examine passages from both the "A" (1781) and "B" (1787) versions of the Transcendental Deduction, as well as remarks contained in both versions of the Paralogisms chapter of the Transcendental Dialectic. I will focus especially on the account of self-consciousness that Kant develops over the course of five paragraphs in §§15-16 of the B-edition Transcendental Deduction. It is there that (i) we find some of Kant's clearest statements concerning the nature of self-consciousness, and (ii) the relationship of self-consciousness to our power of discursive cognition is given prominence.

In both versions of the Transcendental Deduction, Kant describes the cognitive origins of a particular mode of self-consciousness: consciousness of myself as a thinking subject. Kant characterizes this self-consciousness alternatively as "pure apperception" and as a "synthetic unity of apperception." But despite the great importance of this doctrine to the larger aims of the deduction, Kant's remarks on the relationship of self-consciousness to a notion of "synthetic unity" are conspicuously underdeveloped. I nevertheless believe that they provide an account of self-consciousness worth considering in its own right, independently of Kant's larger philosophical ambitions. I will argue that his main insights are the following. Consciousness of oneself *as a subject* cannot be understood on an empirical-observational model, one according to which the "self" of which one is conscious is materially "given" to consciousness. Instead, on my reading, such self-consciousness consists in a distinctive form of awareness of the subject's *activity* of relating together diverse representations in thought—the activity Kant calls "combination."[1] In particular, Kant

[1] Throughout this chapter and subsequent dissertation, I use the terms 'consciousness' and 'awareness' interchangeably.

conceives of this form of self-consciousness as the awareness of oneself as a kind of *agent*. This account reflects his commitment to a conception of thinking as an activity, and to a conception of the subject as an agent of thought.

We can think of this doctrine as providing, in its broad contours, an answer to Hume's skeptical attack against the possibility of being conscious of a "self."[2] By identifying a mode of self-consciousness with the consciousness of one's own *thinking activity*, Kant provides a model for how self-consciousness can be possible without its being construed as the consciousness of some given content. It was the lack of such content that led Hume to cast doubt on the very intelligibility of our concept of a self, which he thought was a "confusion and mistake."[3] Because there is no outwardly or introspectively perceivable content that meets the condition of being identical to the subject, Hume concluded that the self is a fiction. But Kant, as we shall see, not only shares the belief that there is no representational content that is identical to the subject, but shows this conclusion to be necessary, inasmuch as all representational content would have to be relatable to the subject as the self-identical thinker of that content. Once this feature of self-consciousness is presented in its full generality, it can seem rather puzzling how consciousness of a self is so much as possible. Kant's solution to this puzzle is to construe self-consciousness as a form of "agential" consciousness, in a sense to be explained.

I will begin with a concise statement concerning what I take pure apperception to be, followed by some points of clarification. This is in order to anticipate possible sources of confusion and resolve certain terminological ambiguities found in Kant's claims about this mode of self-consciousness. In doing so I shall invoke concepts and terms that will receive fuller explanations in subsequent sections.

---

[2] Hume characterizes his argument as militating against the notion of *personal identity*, because it casts into doubt the very notion of one's "self." But Hume arrives at his skepticism by holding that self-consciousness would require the perception of a simple entity that is identical through time (*A Treatise of Human Nature*, pp. 164-165). Among scholars, there is disagreement over whether Kant was aware of Hume's argument. Patricia Kitcher mounts a compelling case that Kant was indeed familiar with this argument, and provides a brief overview of the scholarly debate up until that point (*Kant's Transcendental Psychology*, pp. 91-100).

[3] *A Treatise of Human Nature*, p. 166.

(**PURE APPERCEPTION**)

Consciousness of myself *as subject* consists in my consciousness of engaging in the activity of combination, where (i) this consciousness is constitutive of that very activity, and (ii) this consciousness constitutively identifies me as the "agent" of that activity.

1. Kant uses different terms to describe this mode of self-consciousness. He calls it "transcendental apperception" (KrV A107), the "transcendental unity of apperception" (A108), the "synthetic unity of apperception" (B131; B136-137; B157), "pure apperception" (B132), and "original apperception" (B132), along with other descriptions. But Kant sometimes uses these and similar terms to denote not self-consciousness *per se*, but rather the relation among our representations in virtue of which such representations can figure as the contents of self-conscious judgments. He also sometimes uses these and similar terms to describe an "objective unity," that is, a relation that holds among our representations insofar as those representations have objective significance.[4] Even in context it is not always clear exactly which of these meanings Kant has in mind.[5] For this reason, unless I am specifically remarking on the relationship of self-consciousness to "synthetic unity," I will use the term 'pure apperception' to denote consciousness of myself *as subject.*
2. Following Kant, the above account of pure apperception defines a mode of self-consciousness in terms of a capacity for consciousness (A108; B132-133). In doing so, it makes an implicit distinction between self-consciousness and consciousness more generally. This raises the question of what Kant means by 'consciousness' [*Bewusstsein*] and how he distinguishes it from *self-*

[4] Cf. B139.

[5] Cf. A108-109 and B132-135. The terminological imprecision partly owes to the fact that Kant ultimately identifies the unity in virtue of which distinct representations belong to a single consciousness and can be represented as such in an act of pure apperception with the unity in virtue of which distinct representations can have, and be recognized as having, objective significance (A108; B137).

consciousness [*Selbstbewusstsein*]. Unfortunately, Kant seldom seems concerned about this distinction, a matter underscored by the fact that he will occasionally use the term 'consciousness' to refer to *self*-consciousness.[6] But the *Jäsche Logic* provides some insight into our question. There Kant asserts that consciousness is the feature of cognition through which cognition is related *to the subject* (JL 9:33).[7] Later, Kant provides a taxonomy of different "degrees" of representation, one in which representation "with consciousness" is placed below the concept-involving form of representation that Kant takes to be a necessary concomitant of self-consciousness (JL 9:64-65). This suggests that consciousness is that faculty through which cognitions are "present to" a subject, which in turn suggests that consciousness is, in contemporary terms, something like the "subjective character" of experience.[8] But this should not be taken to imply that consciousness always consists in some "felt," sensible quality. As I shall argue, pure apperception is our consciousness of ourselves as active thinkers, and no sensible feature of experience could correspond to such consciousness.[9]

3. In saying that pure apperception is a form of self-consciousness that is constitutive of the very activity of which it is a consciousness, I mean to highlight the fact that pure apperception is not a higher-order consciousness, i.e., a representation that is distinct from the relevant activity and takes that activity as its object.

---

[6] Cf. KrV A103.

[7] Cf. KrV A116, A346/B404.

[8] Here I mean 'experience' in a broader, more colloquial sense than Kant's technical notion of experience as "empirical cognition" (A176/B218). Kant himself sometimes uses the term in a broader sense, as when he refers to an "inner experience" of thinking (A343/B401). Cf. Longuenesse (*I, Me, Mine*, p. 88).

[9] I am indebted to Matthew Boyle, whose essay "Kant on Consciousness and the Possibility of Cognition" (unpublished) greatly influenced my thinking on how Kant conceives of the distinction between consciousness and self-consciousness, and which provides a much more thorough investigation into this matter than I can provide here.

4. The phrase 'as subject' [*als Subject*] is employed by Kant in the B Paralogisms chapter to distinguish the kind of self-consciousness distinctive of pure apperception from the kind in which I relate to myself as an object of representation (B407; B410-411). I claim that to be conscious of oneself "as subject" is to be conscious of oneself "as the agent" of thinking activity. The "as X" formulation of such phrases might be taken to suggest that in acts of pure apperception I deploy the concept of a subject or agent. Indeed, in the shared opening paragraphs of the A and B Paralogisms chapters, Kant does refer to the representation "I" as a concept (A342/B400). This seems to imply that representing oneself *as subject* is representing oneself under the *concept* "I." However, to the extent that "I" is a concept at all, it is a concept of a very peculiar sort. Our consciousness of ourselves *as subject*, as an "I," is our consciousness only of the unity of our thinking, *regardless of the content of that thinking*, and Kant takes such consciousness to be a necessary condition of conceptual representation more generally (A341/B399). Thus, "I" is not a representation with a determinate content, and *a fortiori* is not a representation whose content determines its extension.[10] For this reason, later in those same opening paragraphs Kant ultimately denies that "I" is a concept, calling it instead a "wholly empty representation" that expresses "a mere consciousness that accompanies every concept" (A345-346/B404).[11] In this chapter I will attempt to elucidate Kant's notion of pure apperception in terms of a consciousness of oneself as the "agent" of thinking. But for the reasons stated above, this should not be taken to imply that acts of pure apperception deploy the *concept* of agency. Rather, I advert to the phrase 'as the agent' because it captures the relevant sense in which my pure apperception is the consciousness of an activity. Pure apperception is not a consciousness of "agentless" activity (whatever that might be). It is also not a consciousness that

[10] Cf. Longuenesse (*I, Me, Mine*, p. 167 fn. 6).

[11] Strictly speaking, in the cited passage Kant does not claim that "I" *expresses* such consciousness. Rather, he identifies it with such consciousness. But this almost certainly reflects terminological imprecision on Kant's part. Cf. B132, where Kant likewise seems to identify the representation "I think" with the pure apperception that it represents.

something is engaged in an activity, where I happen to be identical to the relevant "something" but that identity does not figure in my awareness. Instead, it is a consciousness that *I* am engaged in that activity, where 'I' identifies me as the thinker, the entity that "does" the thinking. Thus, one might say that pure apperception is a consciousness of *myself doing the thinking*. Alternatively, one might say that pure consciousness is a consciousness of *being engaged in the activity of thinking*. This is the sense in which in pure apperception I am conscious of myself *as the agent* of thinking.[12]

These preliminary comments will become clearer as we proceed. The plan for this chapter is as follows. In §1.1, I provide some necessary background on the project of the *Critique of Pure Reason*. In §1.2, I discuss Kant's principle of the analytical unity of apperception, specifically the representation "I think" and its connection to objective cognition and self-consciousness. In §1.3, I introduce the first of two explanatory challenges that Kant's doctrine of pure apperception is invoked to solve: to account for the possibility of the representation of the "unity" of an object. In §1.4, I introduce the second of those challenges: to account for the self-consciousness expressed by the representation "I think." Finally, in §1.5, I demonstrate how an "agential" conception of self-consciousness is invoked by Kant to meet the explanatory challenges of the preceding two sections. In doing so I aim not to provide a thorough exegesis, much less a defense, of the greater argument of the Transcendental Deduction, but simply to set forth the doctrine of pure apperception that resides within it.

---

[12] Here I join other commentators who interpret pure apperception to be a certain kind of representation of the subject's activity. Cf. Boyle ("Kant on Consciousness and the Possibility of Cognition" (unpublished)), Bristow (*Hegel and the Transformation of Philosophical Critique*, p. 31), Hurley ("Kant on Spontaneity and the Myth of the Giving," p. 145), Longuenesse (*Kant and the Capacity to Judge*, pp. 51-53; *I, Me, Mine*, pp. 6-7, 27, 80-81, 86-87, 107-109), and Smit ("The Role of Reflection in Kant's *Critique of Pure Reason*," pp. 207-208, 220 fn. 5). Longuenesse in particular opts for phrases that are quite close to my own. She writes, for example, "In Kant's Transcendental Deduction…using 'I' in 'I think' expresses the consciousness, by the subject of the activity of thinking, of the unity of the content of her thoughts, and *thereby* of herself as the agent of that unity, whatever the metaphysical nature of that agent might be" (*I, Me, Mine*, p. 81). Other commentators, in explicit opposition to such an interpretation, insist that pure apperception is the representation not of an activity but of a unity that is the result of that activity. Cf. Dickerson (*Kant on Representation and Objectivity*, pp. 87, 133-134) and Strawson (*The Bounds of Sense*, pp. 94-96). But I believe (and as the Longuenesse quotation likewise suggests) that these commentators are missing something distinctive about the Kantian position in supposing that the consciousness of a unity among our representations cannot at the same time be the consciousness of an activity that constitutes that unity. I defend such an interpretation in §1.5.

§1.1 I must offer something of a disclaimer at the outset. To say that Kant has a *theory* of self-consciousness runs the risk of distorting Kant's relatively brief remarks on the topic or reading into them more than Kant intended. In neither the Transcendental Deduction nor the greater part of the first *Critique* does Kant devote significant discussion to a standalone "theory of self-consciousness," one that is argued for independently of, and clearly distinguished from, his larger philosophical project. The two brief sections I am primarily concerned with appear at the very beginning of the B Deduction, and function essentially as preparatory remarks for subsequent arguments. But as Kant announces in the preface to the A edition of the first *Critique*, there are "two sides" to the Transcendental Deduction (KrV Axvi). One side, which makes up Kant's "chief end," is the explanation and demonstration of the objective validity of the pure concepts of the understanding, or "categories." The other side concerns the "powers of cognition on which [the understanding] itself rests," i.e., the understanding "consider[ed]…in a subjective relation" (Axvi-xvii). Within the A Deduction, the "subjective" component consists of an exposition of three different powers of "synthesis," which correspond, in ascending order, with a capacity for spatiotemporally-ordered intuition, empirical association, and conceptual representation, respectively (A95-114).[13] Kant is especially concerned to explain the relationship of our power of conceptual representation to the two lower powers of synthesis, and argues that a capacity for pure apperception is a condition of the former (A103-110). However, this "side" of the Deduction is advanced by Kant only to the extent he thought necessary to demonstrate the objective validity of the categories. And while Kant would eventually completely rewrite the Transcendental Deduction, Kant's chief concern remained the demonstration of the objective validity of the categories. Thus, whatever "theory of self-consciousness" we might be able to extract from Kant's remarks on that topic, it was not Kant's primary goal to articulate and defend it.

[13] Kant labels these powers the "synthesis of apprehension in the intuition" (A98), the "synthesis of reproduction in the imagination" (A100), and the "synthesis of recognition in the concept" (A103). Kant's notion of synthesis will be explained in §1.3.

For this reason, it is impossible to discuss Kant's claims about the nature of self-consciousness without mentioning how that topic figures in the greater project of the Transcendental Deduction. This in turn requires taking a brief foray into Kant's theory of cognition, as well as orienting the project of the Deduction within the more general aims and commitments of the first *Critique*.

One fundamental commitment of the *Critique* is that our representation of objects (and thus our knowledge of objects) involves the cooperation of passive and active representational capacities. This interplay of "receptivity" and "spontaneity" is reflected in Kant's taxonomy of representations of objects, or "cognitions." Cognitions fall into two kinds: intuitions and concepts. An intuition originates in the subject's faculty of sensibility, which is a receptive faculty, "[t]he capacity to acquire representations through the way in which we are affected by objects" (A19/B33).[14] An intuition applies to its object "immediately," in that the intentional link between it and its object is not mediated by some further representation. Moreover, it is a singular representation, applying by its nature to only one object. Concepts, by contrast, are general representations (hence applying to all objects of a sort designated by the concept) by means of which we "think" objects. *Thought* as it is understood here is the specific purview of the understanding, "the faculty for bringing forth representations itself, or the spontaneity of cognition" (A51/B75). Because we apply concepts to objects through acts of judgment that relate concepts to each other, Kant also refers to the understanding as a "faculty for judging" (A69/B94). Concepts apply to their objects "mediately," in that their relation to their object is mediated by the presence of an intuition. This feature explains why discursive subjects—subjects who, like us, represent objects by subsuming them under concepts in acts of judgment—necessarily rely on

[14] That is not to say that an intuition (*Anschauung*) is a sensation (*Empfindung*). Kant defines 'sensation' as a "perception that refers to the subject as a modification of its state" (A320/B376). Sensations are distinguished from cognitions (*Erkenntnisse*) by the fact that they are *mere* modifications of the subject's mental life and hence do not refer to objects. All of these mental states belong to the genus of representation (*Vorstellung*), so Kant seems to include under the label "representation" mental states that do not themselves represent. Cf. George ("Kant's Sensationism"; "*Vorstellung* and *Erkenntnis* in Kant") and Pereboom ("Kant on Intentionality"). For a different take on whether all *Vorstellungen* represent, see Tolley ("Kant on the Content of Cognition").

intuitions for knowledge, and thus stand in a receptive relation to the objects of their knowledge. We see that the discursivity of the human intellect implies that our overall capacity for representation involves both active and passive representational powers.

One of the chief aims of the *Critique of Pure Reason* is to vindicate the possibility of synthetic, *a priori* knowledge. Kant considered the Transcendental Deduction to be the most crucial step in this undertaking (Axvi-xvii). The goal of the Deduction is to explain the possibility of a representation bearing a non-accidental and intentional relation to an empirically-given object, in such a way that it could provide the materials for synthetic, *a priori* knowledge of an external world. In the preceding chapter—the so-called "metaphysical deduction" of the categories (A66-83/B91-116)[15]—Kant argued that there are certain *a priori* concepts, or "categories," related to the basic functions of the understanding in such a way that subjects necessarily apply them whenever they regard something *as an object*. So, for example, to think of something as an object is to think of it in terms of the relation of a substance to its accidents, or as standing in relations of cause and effect. However, even granting the success of this argument, it shows only that the categories are internal conditions on our taking representations to have objective significance. It does not demonstrate that the objects we encounter through sensible intuition in fact stand in categorial relations. That task is left to the Transcendental Deduction, which attempts to show "how subjective conditions of thinking should have objective validity" (A89-90/B122).

There is deep disagreement among commentators about how best to characterize the argument of the Transcendental Deduction.[16] I hope, however, that the following broad

[15] Kant added two sections to this chapter in the B edition. He refers to this chapter as a "metaphysical deduction" at B159.

[16] Such interpretative questions include, but are not limited to: (i) whether (and if so, in what ways) the A and B versions of the Transcendental Deduction amount to distinct arguments, or simply different expressions of what is fundamentally the same argument; (ii) the overall function of the Transcendental Deduction with respect to the *Critique of Pure Reason*, most notably as it pertains to the aims of the prior "metaphysical deduction" and subsequent Analytic of Principles; (iii) how to interpret Kant's principle of the analytical unity of apperception; (iv) whether the Transcendental Deduction

characterization of §§15-20 of the B Deduction will be uncontroversial enough. First, after setting forth his theory of synthesis, Kant claims that in order to represent something in thought it must be possible to represent the identity of the thinker (B131-132). Second, Kant argues that our ability to represent the identity of the thinker presupposes a kind of self-consciousness that he labels 'pure apperception' or the 'synthetic unity of apperception' (B131-136). Finally, Kant argues for a necessary connection between the unity of self-consciousness and our capacity to cognize objects through the categories of the understanding (B136-B143). The upshot is that a subject's ability to make empirical judgments requires that empirical objects are such that the categories necessarily apply to them.[17] Thus, within the greater argumentative framework of the Deduction, Kant attempts to demonstrate the objective validity of the categories by way of advancing a conception of self-consciousness as the representation of a "synthetic unity," in a sense to be explained.

§1.2 In the B Deduction, Kant first explicitly introduces the topic of self-consciousness in §16, which is titled "On the original-synthetic unity of apperception" (B131). He begins by pointing to what he takes to be an analytic connection between the ability to represent content in thought and the ability to represent one's identity as the subject of that thought. According to Kant,

> The **I think** must **be able** to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, which is as much to say that the representation would either be impossible or else at least would be nothing for me. That representation that can be given prior to all thinking is called intuition. Thus all manifold of intuition has a necessary relation to the I think in the same subject in which this manifold is to be encountered. (KrV B131-132)

assumes only a capacity for self-consciousness, or whether in addition it assumes a capacity for empirical cognition; and (v) whether the Deduction entails conceptualism about the content of experience. For a helpful overview of certain ongoing exegetical disputes, see Pereboom ("Kant's Transcendental Arguments").

[17] Somewhat more controversially, I believe that in the remainder of the B Deduction (§§21-27), Kant attempts to show *how* the conclusion of §20 is possible, that is, *how* categories of the understanding are objectively valid. He does this by arguing that space and time, the forms of our intuition of objects, are themselves synthesized in such a way that any object structured by those forms is likewise structured by the categories. Cf. Longuenesse (*Kant and the Capacity to Judge*, pp. 211-233).

The representation "I think" that Kant says can "accompany" my representations is what he subsequently refers to as the "analytical unity of apperception," or the representation of "the identity of the consciousness in these representations" (B133). Following Kant's usage, then, let's call the principle that Kant sets forth in the quoted passage the 'principle of the analytical unity of apperception'. Kant's thought is that for a certain range of representations, the subject must be able to represent herself as the subject of those representations. As the ensuing discussion makes clear, Kant holds this principle to be an analytic truth, known through *a priori* reflection on what is contained in the thought that a certain kind of subject is in a certain kind of conscious representational state (B135; B138).

I use qualifying language like 'certain kind of' because it's far from obvious what the scope of the principle is meant to include. Kant begins by making what seems to be an unrestricted claim about all possible representations a subject might possess: "The I think must be able to accompany <u>all</u> my representations" (B131, my emphasis). However, he immediately proceeds to qualify this remark by suggesting that I could have representations that the representation "I think" could not accompany: such representations might be possible, but they would be "nothing for me" (B132).

Let's suppose—as an initial hypothesis that I will argue is mistaken—that (i) the principle of the analytical unity of apperception expresses a formal condition on the representations of *all* conscious subjects, and (ii) a representation is "something for" a subject (i.e., not "nothing for" a subject) just in case it is a conscious representation. That principle makes an extremely strong claim, one that is clearly not an analytic truth. For that principle states that *any* conscious representation can be recognized by the relevant subject as belonging *to* the subject, and that such recognition can be expressed in a judgment that represents the subject as the thinker "I" in "I think." It therefore denies the possibility of a conscious subject that is not likewise a *self*-conscious subject, indeed a subject that is conscious of itself being engaged in thought. Kant, however, acknowledged the existence of

conscious, non-human animal subjects (JL 9:64-65). Moreover, Kant associated the capacity for self-conscious representation with sophisticated cognitive abilities such as concept deployment and logical inference. Kant therefore could not have plausibly and coherently thought that, necessarily, every conscious being has the capacity for self-consciousness.

One clue for the proper interpretation of the principle is provided by the fact that Kant identifies representations that are not "for" their subject with those that "could not be thought at all" (B132, my emphasis). As we saw, Kant identifies *thinking* with the understanding's capacity to deploy concepts in acts of judgment. This suggests that he restricts the scope of the principle to representations (whether concepts or intuitions) that could serve as part of the *content of judgments*. Yet another clue is provided by the fact that the principle is stated and argued for in the first person: Kant is expecting the reader to consider, in the first person, the conditions under which, e.g., my representations could be represented as *mine* and thus as something for *me*. The principle is therefore likely restricted to discursive subjects, thereby allowing for the possibility of other kinds of subjects in answer to the concerns raised above. (This would also help to explain how Kant can make the unrestricted claim that "all manifold of intuition has a necessary relation to the I think in the same subject in which this manifold is to be encountered" (ibid., my emphasis).) Without fully settling these interpretive questions, we can say that the principle of the analytical unity of apperception entails *at least* the following: for a discursive subject, any representation that could serve as part of the content of that subject's judgment is such that it could be identified by the subject as belonging *to* the subject, the "I" that thinks that content. In the adverbial mode: any such representation could be possessed *self-consciously*.

As was mentioned above, Kant claims that the representation "I think" that can accompany my representations is the representation of the "identity of the consciousness in these representations" (B133, my emphasis). Likewise, in the A Deduction, when introducing the topic of pure apperception,

Kant refers to "a consciousness of the identity of oneself" (A108, my emphasis). Since (as we'll see) the representation "I think" is an expression of pure apperception, such comments imply that the "I" in "I think" represents the identity of the thinking subject. But here we should be careful, for Kant makes it clear in both the A and B versions of the Paralogisms chapter that the representation "I think," which Kant calls "the sole text of rational psychology," does not express the consciousness of a simple substance that persists through time (A343/B402). In connection to the third paralogism, which concerns personal identity, Kant states that "the identity of the consciousness of Myself in different times is…only a formal condition of my thoughts and their connection, and does not prove at all the numerical identity of my subject" (A363).[18] In other words, the mere representation "I think" provides no insight into the real nature of the referent of "I." In light of this remark, we should interpret the principle of the analytical unity of apperception only in terms of a formal condition on our representations. That condition includes the possibility of the representation of my diachronic identity: for any two representations, including representations at different times, insofar as those representations can figure as the content of my judgments, it must be possible to represent *myself* as the self-identical thinker of those representations. The representation of myself as the diachronically selfsame thinker of my representations is thus a condition on those very representations, regardless of my ultimate metaphysical nature.[19]

Kant therefore supposes that there is an analytic connection between the possibility of the representation of the identity of the subject ("I think") and the possibility of discursive representation

[18] Likewise, in the B Paralogisms chapter Kant claims that the consciousness of the identity of oneself expressed by "I think" is not a subject's "consciousness of the identity of its own substance as a thinking being in all changes of state" (B408).

[19] The general theme of the A and B Paralogisms chapters is that formal conditions on representation provide no basis for inferring a metaphysical doctrine of the self. Cf. Longuenesse (*I, Me, Mine,* pp. 102-169). To forestall misunderstanding, it must be stated that what we are after in this chapter is neither (i) a metaphysical doctrine of the self nor (ii) an account of the ultimate noumenal grounds of our capacity for self-conscious representation. Any such doctrine would conflict with the anti-metaphysical spirit of the first *Critique*.

more generally. The "I think," as a representation of the identity of the subject, expresses a kind of self-consciousness: the consciousness that my representations are *mine* (B132). The principle of the analytical unity of apperception thus specifies a necessary condition on all representations that are "mine" in the sense stipulated above: "as my representations (even if I am not conscious of them as such) they must yet necessarily be in accord with the condition under which alone they **can** stand together in a universal self-consciousness, because otherwise they would not throughout belong to me" (B132-133).

Kant refers to the relationship such representations bear to one another as an "original combination" (B133), or, alternatively, an "original synthetic unity" (B131; B135; B137). Kant's primary aim is to show that the conditions of this "synthetic unity" function as a constraint on (and perhaps ground of) the objects given by a discursive subject's power of sensible intuition. Kant will argue that the applicability of the categories to the objects of our sensible intuition is a condition on the possibility of our being able to represent such objects in thought. For this reason, commentary has tended to focus on whether this argument succeeds. But I would like to set this question aside and focus instead on the *conception* of self-consciousness that Kant advances. What is the nature of the self-consciousness that Kant claims is a necessary concomitant of discursive representation? What is involved in the recognition that my representations are *my* representations?

§1.3 I will argue that Kant conceives of self-consciousness as a subject's consciousness of being engaged in the activity he calls "combination." On my interpretation, Kant's account of self-consciousness serves as a solution to two explanatory challenges that he takes to be intrinsically related. The first is to show how I can represent the *unity* (in a sense to be explained) that any manifold of representations must possess insofar as such representations are of a single object, despite the fact that my representations are not *given* to me as unified. The second is to show how I can represent my

own unity, as the self-identical subject of these representations, despite the fact that my relationship to these representations as their subject does not figure in the content of the representations themselves.

To appreciate the first problem, we must look to Kant's doctrine of synthesis. What follows might appear to be unrelated to the topic of self-consciousness, but its relevance to that topic will become clearer in due course. To begin, consider the following representations:

(i) a sensible intuition of *redness* and *heat* in an object

(ii) a judgment that an object given by a sensible intuition is red and hot

(iii) a judgment that some red, hot object is a stone

(iv) a judgment that the sun causes the stone to be red and hot

(v) a judgment that the stone is not edible

(vi) a judgment that, necessarily, no stone is edible

I present this list of possible representations to provide a sense of what Kant means by a representation's "unity." These representations are alike in that each is a distinct, individual representation. For example, an intuition of redness and heat in an object is, despite its complexity, nonetheless a *single* intuition. But each of the above representations is constituted out of other, numerically distinct representations in a way that expresses some relation—in Kant's idiom, a certain "unity"—contained in or exhibited by the object or objects of representation. For example, the causal relation that the sun bears to the stone is expressed by the thought, in (iv), that the sun causes the stone to be red and hot. That thought relates together the concepts [sun], [stone], [red], and [hot] in a way that purports to have objective significance, that is, in a way that expresses a certain relation (unity) of objects and their features. Moreover, the thought itself contains (or, better, *is*) a certain unity of

concepts. It (the thought, the judgment) is *one representation* constituted out of the various representations (concepts) it deploys. It is a *unified* representation.[20]

Kant addresses the question of how we are able to represent the unity exhibited in an object in the opening section (§15) of the B Deduction. This section immediately precedes the one we just discussed (§16), which we said opens with a statement of the principle of the analytical unity of apperception. Kant begins §15 by noting that passive affection through sensibility supplies us with a diverse multiplicity (or "manifold") of representations. But according to Kant, mere sensibility does not establish any unity among the given representations; in other words, sensibility does not suffice for the generation of *unified* representations. Rather, unity among our representations originates in an active power that he alternately labels 'combination' and 'synthesis':

> the **combination** (*conjunctio*) of a manifold in general can never come to us through the senses . . . [A]ll combination, whether we are conscious of it or not, whether it is a combination of the manifold of intuition or of several concepts…is an action of the understanding, which we would designate with the general title **synthesis** in order at the same time to draw attention to the fact that we can represent nothing as combined in the object without having previously combined it ourselves. (B130, my underline)

Kant makes no attempt to defend the claim that unity among our representations cannot arise through sensibility and therefore originates in an operation of the understanding. Moreover, it is not clear why

[20] Two points of clarification are in order. First, a question arises as to whether by 'representation' (as well as terms like 'thought' and 'judgment') I mean the act of *representing* or something like the abstract mental item that stands in an intentional relation to an empirical object and is the content, or outcome, of such an act. (Compare [perceiving X] and [having a perception of X].) For now, I mean the latter, but only because I don't want to prejudge the question of the role of our combining activity (and our consciousness thereof) in generating representations in the latter sense. As I will argue in §1.5, on Kant's account such representations of unity are possible only in virtue of the self-conscious nature of our representational activity. Second, my examples may give the impression that a single-clause judgment is a representation of unity only when it involves multiple token predicates. But this is not so. Consider, for example, an occurrent representation of my desk *as a desk* in the ostensive judgment *this is a desk*. Such a representation consists in my relating and thereby uniting various of my sensible impressions as impressions of a single object; this is what gives me a single, unified intuition of the object before me. Furthermore, my applying the concept *desk* to this object involves my identifying various of its features—its flat, smooth surface; its four legs—and relating them to other possible experiences in which I might encounter an object that shares these features and belongs to the kind "desk." So the relating involved in conceptualizing the object as a desk is also a unifying inasmuch as the relating involves the identification of commonalities across various representations. That is, the identification of these commonalities is such that when I represent this object as a desk I do not just relate to my immediate intuition of the object before me but all intuitions of a certain kind; and this feature of my representation makes my deployment of the concept *desk* a representation of unity, in this case, the unity of these various possible intuitions.

Kant is entitled to this premise[21]—a point I shall return to in my discussion of Kant's methodology in the next chapter. For now, note the underlined portion of the passage, which makes explicit a corollary of this premise, namely, that our ability to represent the objective unity of diverse phenomena—as he puts it, to represent diverse characteristics "as combined in" an object—requires an act of synthesis.

Now, to represent diverse features of an object *as combined* is to represent the *unity* of the object, the relation of its various features in the constitution of the object, as in the above judgment [(ii)] that a given object is red and hot. This reflects the fact that synthesis is not the representation of *simplicity,* but instead the "holding together" of a *multiplicity*.[22] Furthermore, the representation of unity makes possible a discursive *analysis* of this very unity: my judgment that a given object is red and hot can serve as a basis for my distinguishing the object's redness from its hotness.[23] As Kant remarks, "the dissolution (analysis) that seems to be [the] opposite [of synthesis], in fact always presupposes [synthesis]" (B130). But note that in order for my representation of unity to serve as a basis for analysis, I must recognize not just the unity of the object of which it is a representation but the unity of the representation itself.[24] For to analyze a representation is to regard it—that very representation—in

[21] Cf. Guyer, who identifies this as a premise of the argument of the Transcendental Deduction ("Kant on Apperception and 'A Priori' Synthesis," p. 207).

[22] Cf. Longuenesse (*Kant and the Capacity to Judge*, p. 38 fn. 10).

[23] In saying that a judgment that predicates two properties of a single object is a representation that can serve as an analysis of that object in terms of those properties, I do not mean to deny that the judgment itself already distinguishes between those two properties. To the contrary, that is exactly my point. If the judgment did not already distinguish between such properties—if in making the judgment I were not aware of the unity of the object and therefore of the relation expressed by the judgment—then my judgment could not itself serve as the basis of a discursive analysis of the object, which in our example would be expressed by two judgments: (i) the object is red, and (ii) the object is hot. The fact that the original judgment *can* serve as the basis for such an analysis entails that it—the very judgment—is a representation of a certain kind of unity. I also do not mean to deny that the judgment already expresses an analysis of the sensible manifold. In *Kant and the Capacity to Judge*, Beatrice Longuenesse argues that there is a necessary synthesis of the sensible manifold that makes possible an *analysis* of that very manifold, guided by the logical forms of judgment, that "reflects" the synthetic unity of the manifold under concepts in particular acts of judgment. If that is right, then the empirical judgment that some object is red and hot is made possible by a prior analysis of the sensible manifold. But that is fully compatible with my claim.

[24] Cf. Rödl ("Self-Consciousness and Knowledge," p. 362).

respect of its constituent elements. That these constituent elements are related together in the first place reflects their being "held together" or synthesized in a representational act.

However, while the synthesis of a manifold is necessary for the representation of a unity *as a unity*—e.g., not just the representation of several features combined in an object, but the representation of them *as combined* in the object—it is not sufficient. In order to explain why, I will now introduce a terminological distinction that is suggested by Kant's own remarks. (That said, it is not completely clear that the distinction is Kant's own, and at any rate Kant does not maintain the distinction with perfect consistency.) Earlier I claimed that 'synthesis' and 'combination' are synonymous labels for a single type of mental act, the act of bringing representations together to constitute a relation, or unity, among them. Going forward, I will use the term 'synthesis' for the general power of bringing representations together,[25] and I will use the term 'combination' for the exercise of that power whereby the subject is also conscious *of the unity* constituted by that power.[26] Thus, combination is a special kind of synthesis, the kind of synthesis that does not just bring about a unity but in bringing about that unity represents the very unity that it brings about.

I claim that the model of synthesis—the metaphor of a "holding together" of a diverse manifold of discrete representations—does not suffice for a subject's representation of the unity of the unified representation that results. The point is a simple one: to impose a relation (unity) on a manifold through an act of synthesis is not necessarily to represent the relation constituted by that act *as* the relation thereby constituted. And despite the fact that Kant is not always careful to distinguish mere synthesis from the power I call 'combination', Kant is sensitive to this distinction. For example, when Kant introduces the notion of synthesis in the preceding "metaphysical deduction," he refers to

[25] Cf. KrV B130, where Kant refers to 'synthesis' as a "general title" of the representational power.

[26] Cf. B130-131, where Kant claims that combination "is the representation of the **synthetic** unity of the manifold."

a prediscursive type of synthesis in which, presumably, no representation of unity *as unity* is yet included:

> The synthesis of a manifold…first brings forth a cognition, which to be sure may initially be raw and confused, and thus in need of analysis; yet the synthesis alone is that which properly collects the elements for cognitions and unifies them into a certain content; it is therefore the first thing to which we have to attend if we wish to judge about the first origin of our cognition.
>
> Synthesis in general is, as we shall subsequently see, the mere effect of the imagination, of a blind though indispensable function of the soul, without which we would have no cognition at all, but of which we are seldom even conscious. Yet to bring this synthesis **to concepts** is a function that pertains to the understanding, and by means of which it first provides cognition in the proper sense. (A77/B103-A78/B103).

Kant speaks here of a mental process *of which we are not even conscious* that unifies representations into a certain content, the outcome of which is a "raw and confused" concatenation of representations that is "in need of analysis" (ibid.). I submit that by such a process of synthesis Kant must mean a prediscursive synthesis that constitutes a certain relation among our representations, but that does not (yet) include a representation *of the relation* thereby constituted. A candidate for such a synthesis is the spatiotemporal ordering of sensations of redness and heat in an intuition of an object. Merely to have such a unified intuition is not yet to represent the unity of that intuition by subsuming it under the concepts of redness and heat in an empirical judgment. In doing so we would "bring this synthesis to concepts" (ibid.).[27]

For this reason, the notion of synthesis given thus far doesn't ultimately explain how I can represent the unity of an object. I believe that Kant has precisely this point in mind in a remark he makes near the end of §15:

> But in addition to the concept of the manifold and of its synthesis, the concept of combination also carries with it the concept of the unity of the manifold. Combination is the representation of the **synthetic** unity of the manifold. The representation of this unity cannot, therefore, arise from the combination; rather, by being added to the representation of the manifold, it first makes the concept of combination possible. (B131)

[27] See footnotes 20 and 23 for an account of how this might come about.

This is an obscure passage, not least because at first glance it's difficult to reconcile (i) the claim that combination *just is* the representation of the synthetic unity of the manifold, with (ii) the claim that the representation of such unity cannot "arise" from the combination. But I think sense can be made of it if we attend to the insufficiency of "mere" synthesis in accounting for our ability to represent the unity of an object.

To perform an act of combination is to relate diverse representations in a certain way, i.e., to bring about a certain unity among a manifold of representations. That is why "the concept of the manifold and of its synthesis" is included in the concept of combination (ibid.). But Kant proceeds to make a further claim that is not entailed by this one and for which he offers no argument. He asserts that combination "is the representation of the synthetic unity of the manifold," i.e., is the representation of the unity not just of the represented object, but of the manifold of constituent representations so unified. However, even granting the above argument that in order to represent the unity of an object we need to be able to represent the unity *of the unified representation*, it is not clear why the unity of a token unified representation must be represented by *that very token representation* rather than a distinct, second-order representation that takes that first-order representation (and its unity) for *its* object.

Another worry about this passage is that, on a certain reading at least, Kant seems to retract this claim in the very next sentence. He writes, "The representation of this unity cannot, therefore, arise from the combination; rather, by being added to the representation of the manifold, it first makes the concept of combination possible" (B131). But if the act of combination *is* the representation of unity, in what sense does the representation of unity not "arise" from this act? Furthermore, Kant's claim that the representation of unity must be "added" to the manifold might be taken to suggest that we "add" this representation by means of a distinct, second-order representation.

In response to the first concern, it is worth stepping back and reflecting on the nature of our investigation. We are inquiring into how subjects can, on the basis of a single representation, represent the unity of an object. A presupposition of this inquiry is the *sufficiency* of the representation to its task. We do not explain that sufficiency by positing a *separate* representation. More importantly, by invoking a second-order representation we do not *explain* how subjects are able to represent the unity of a diverse manifold, but in fact only presuppose that they can. This is because the above argument concerning the analysis of unified representations can be applied *to the second-order representation*. I claimed that in order to represent the unity of an object, the subject must be able to represent the unity of that object's very representation. But on the proposal under consideration, that representation is itself an object of a second-order representation. Our grasp of the complexity of the first-order representation thus requires a grasp of the unity of the *second*-order representation—a grasp of how *its* constituent elements are related. And so on for representations of still higher orders. Therefore, appeals to higher orders of representation cannot explain how subjects are able to represent the unity of an object. I believe it was on these grounds that Kant held that acts of combination must include a representation of the unity of the very manifold of representations so unified.

How, then, should we understand Kant's assertion that the representation of unity cannot "arise from" [*entstehen*] the combination? I want to suggest that the apparent inconsistency of this claim with my interpretation reflects an ambiguity in Kant's use of the term 'combination' [*Verbindung*]. '*Verbindung*' can refer either to (i) the *act* of combining a manifold of representations in such a way that we represent the unity achieved through that act, or to (ii) the *state* of the combined manifold.[28] When Kant says that combination includes a representation of the unity of the synthesized manifold, he is referring to the representational act whereby a manifold of discrete representations is unified and how,

[28] The use of '*Verbindung*' to denote a state occurs elsewhere in the same paragraph. Kant claims, e.g., that when we judge according to certain logical functions, "combination, thus the unity of given concepts, is already thought" (B131). It seems relatively clear that what is "thought" in these judgments is the unity or combination of its concepts.

through that very act, we are conscious of its unity. But when he claims, as he does in the quoted passage, that our representation of this unity cannot "arise from" the combination, I believe he is referring to the manifold's combined state. I take his point to be an expression of the point I made above. Merely to have a multiplicity of representations that stand in a certain relation is not to have a representation of the relation in which those representations stand. Furthermore, the unity of the manifold cannot, so to speak, be "read off" the various discrete representations that constitute the manifold, because the unity of the manifold is *not given or contained in the content of those representations*. Thus, when Kant claims that a representation of unity needs to be "added" to the manifold, he is not invoking the need for a second-order representation but pointing to the fact that the unity of the manifold is not included in the manifold's content.

That is why the metaphor of a subject's "holding together" diverse representations, taken in isolation, is inadequate to the task of accounting for our ability to represent the unity of an object. The fact that a subject can impose a certain relation on her representations is not sufficient to explain how she is able to recognize that they stand in such a relation. What seems to be missing from the model of synthesis given thus far is any awareness on the part of the subject concerning *how* her bringing or holding those representations together constitutes their unity in a single representation. On this basis, Kant signposts the introduction of a new topic, claiming that "[w]e must therefore seek this unity…in that which itself contains the ground of the unity of different concepts in judgment" (B131). This is an implicit reference to the synthetic unity of apperception, the formal unity in virtue of which my various representations belong, and can be represented as belonging, to me as their self-identical subject.

Kant's complete explanation of our ability to represent the unity of an object invokes the categories of the understanding, as concepts that correspond to the "necessary synthetic unity" of our intuitions insofar as such intuitions can figure as the content of self-conscious judgments. This

component of the Transcendental Deduction would take us well beyond the scope of this chapter. Nevertheless, Kant's reference to the synthetic unity of apperception at the end of §15 suggests that the conception of self-consciousness as a representation of a "synthetic unity," which Kant first introduces in §16, will serve in part to *complete* the account of how subjects represent the unity of an object. In arguing for the inadequacy of "mere" synthesis in accounting for the representation of unity, I cited the fact that the unity of a given manifold is not included in the manifold's content. I argue in the next section that a formally identical difficulty arises for the explanation of the self-consciousness expressed by "I think." To understand the account of self-consciousness that, on my interpretation, Kant provides, we must appreciate its role in overcoming these explanatory hurdles.

§1.4 Section §1.2 ended on a question: what is the nature of the self-consciousness expressed by "I think"? There we said that the applicability of the representation "I think" is a condition on every representation that could figure in the content of a judgment. Such a conclusion is entailed by the principle of the analytical unity of apperception.

After setting forth this principle, Kant asserts that the representation "I think" expresses and is made possible by "pure" apperception, which he contrasts with "empirical" apperception (B132). Empirical apperception is a form of self-awareness grounded in what Kant calls "inner sense," the mind's capacity to "intuit…its inner state" (A22/B37).[29] Inner sense encompasses representations of how "we appear to ourselves" insofar as we are "internally affected" (B153). The relevant point for our purposes is that inner sense is a form of receptivity and thus operates according to an empirical-observational paradigm. Through inner sense we are a kind of spectator to the modifications of our own minds, as, for example, when we observe that we are having a perception of the color blue. The

[29] This point is supported by the fact that later in the B Deduction Kant explicitly contrasts inner sense with the "faculty of apperception" (B153). It is clear from the surrounding discussion that Kant has in mind *pure* apperception.

representations of inner sense take as their "objects" the subject's own representations. That is why inner sense makes possible an empirical *apperception*, for it is the subject's empirical consciousness of herself. But Kant insists that the representation "I think" is not an expression of any sort of empirical consciousness. As he says later in the B Deduction, the representation "I think" captures my relation to myself "as intelligence and thinking subject" and not "as an object that is thought" (B155).[30] Kant therefore holds that the self-consciousness represented by "I think" cannot be understood on an empirical-observational model, one according to which we represent the "self" as an object *given* to consciousness.

The inadequacy of empirical apperception to the explanation of the representation "I think" parallels a point made in the previous section. Just as the unity of a manifold is not included in that manifold's content, so the self-consciousness of which "I think" is an expression cannot take the form of an awareness of some given content, such as the content of inner sense. To see why, notice that "I think" expresses a relation, or unity, that obtains between the representations that constitute a judgment, in exactly the sense of 'unity' adduced above. This is because token deployments of "I think" assert of some manifold of representations that all of its representations are related to me, their subject, *as* the subject of those representations. When I judge, for example, that *I think the sun caused the stone to be red and hot*, I not only relate the relevant constituent representations to each other in such a way as to reflect an objective relation that obtains between the sun, the stone, and certain of the stone's properties; I moreover expressly relate the constituent representations to *me* as the subject of those representations, expressing my recognition that they are together *my* representations. But given the lessons of the previous section, this yields the conclusion that my basis for regarding my representations as in the relevant sense *mine* is not to be found in the content of the representations themselves. For just as, in general, the unity of a given manifold is not given or included in the content

[30] Cf. B158-159, B407, B429ff.

of that manifold, so the kind of unity expressed by "I think" is not given in the content of the relevant manifold.

And just as before, we cannot overcome this difficulty by appealing to second-order representations. This explains why "I think" cannot be construed as an expression of empirical apperception. We said that empirical apperception consists of higher-order representations that take as their objects particular modifications of our minds. To make this concrete, suppose there is a second-order representation whose object is a particular unified manifold of perceptual content, e.g., a sustained visual representation of the color blue. There is nothing in this description that accounts for my ability to represent the visual representation *as* belonging to my subjectivity. That is not to deny that I have such an ability; indeed, according to the principle of the analytical unity of apperception, I am able to represent my identity as the subject of this visual representation in the ostensive judgment *I think that is a visual representation of the color blue.*[31] In this example, we represent not just the existence of a particular visual representation, but also the relation of this representation to me as its subject. But the question we are trying to answer is how representations of this sort are possible; as before, appealing to second-order representations does not *explain* how we are able to represent unity—in this case, the unity that representations possess insofar as they are together *mine*—but only presupposes that we can. In particular, in the example just given, "I think" does not represent my empirical apperception of the visual representation but instead the consciousness of myself as the *subject of* the empirical apperception. (As Kant might put this point, it represents the pure apperception that accompanies every empirical apperception.) Kant glosses this in the claim that "the empirical consciousness that accompanies different representations is by itself dispersed and without relation to the identity of the subject" (B133, my emphasis). I take the qualification 'by itself' to be significant: Kant's point is not, of course, that empirical apperception bears no relation to the "I think" but rather

[31] Here I assume that my second-order representation can figure in the content of a judgment.

that there is nothing in the nature of empirical apperception—empirical apperception taken "by itself," as it were—that accounts for its connection to a subject and our ability to represent this connection.

The discussion so far has been pitched at a high level of abstraction, but its main implications can be seen through a more direct appreciation of what it is to represent ourselves as a thinking subject. When I judge that *I think* something, it is not because I've turned my mind's eye inward and located some ego that I identify with my subjectivity. More generally, pure apperception is a form of self-consciousness that, unlike empirical apperception, cannot itself contribute any *additional* representational content to consciousness. Kant summarily expresses this thought in the claim that pure apperception "is that self-consciousness which, because it produces the representation *I think*, which must be able to accompany all others and which in all consciousness is one and the same, cannot be accompanied by any further representation" (B132). Commentator William Bristow helps to unpack this claim:

> All my representations must allow of being made self-conscious in the sense that each must allow of being a substitution instance of x in the thought 'I think x'. But in the sense in which we say this of all other representations, we cannot say it of the representation 'I think' itself, exactly because this representation expresses our subjectivity. Kant claims that all representations bear a necessary relation to the 'I think' in the same subject in which they occur. But insofar as [the analytical unity of] apperception, the 'I think', itself expresses *the necessary relating* of all cognitive representations to the identical subject, to me as the self-identical thinker of them, the 'I think' cannot itself be present to me as a particular content of representation. No particular representational content could express the 'mine-ness' (the necessary relation to me) of that content, in this sense, since *any* particular cognitive content (even if it were to represent me in some sense) must be again related to me... In short, no representational content could express the 'I' *as subject*.[32]

In short, pure apperception is a form of self-consciousness that cannot take the form of an awareness of given content. The content of my subjective awareness is the particular mode of presentation of the object I represent; when I am conscious of the relation of my representation to me as its subject—when I represent the object *self-consciously*—I distinguish that content from my subjectivity. That is, to

[32] *Hegel and the Transformation of Philosophical Critique*, p. 30.

represent some given content as (in the relevant sense) *mine* involves representing an 'I' as distinct from that content, for any possible content; and just because of that, the mode of presentation for the 'I' cannot take the form of additional content.

But that is just to say what self-consciousness *isn't,* not what it is. I have argued that an account of self-consciousness faces difficulties that are formally identical to those we encountered in the previous section, where we considered the subject's capacity to represent the unity of an object. I take up Kant's positive account of self-consciousness in the next section. There I will explain how it serves as a common solution to these problems.

§1.5 Each of the previous two sections presents us with a puzzle. First, how do we represent a diversity of features as "combined" in an object? Second, how do we represent ourselves as the subject of such representations? The theme that unites these sections is that both are representations of unity, in particular, representations of the unity of a diverse manifold of representations.

In each case the relevant explanatory hurdle issues from a single premise, namely, that the unity of a manifold cannot be given by, or contained in, the content of the manifold itself. This hurdle is only partially overcome by the doctrine of synthesis Kant sets forth in §15. He claims that, since the unity of a manifold is not given in the content of the manifold, any representation of unity requires an act of synthesis on the part of the subject. But as we noted, imposing a certain unity on the manifold does not entail being conscious of the unity that results. Without an account of how the subject recognizes the unity constituted by the synthetic act, the explanation is at best incomplete. The explanatory challenges of the last two sections are thus two sides of a common coin.

Notice, however, that if (i) all representation of unity requires an act of synthesis that brings about that very unity and (ii) pure apperception involves a representation of unity, then pure apperception will likewise require, and be made possible by, synthesis. Kant alludes to this shortly

after introducing the principle of the analytical unity of apperception, in his remark that the representation "I think" requires "an act of spontaneity" (B132). Kant is referring to the spontaneity evinced in acts of synthesis. With this, we begin to see why Kant identifies pure apperception with a representation of a "synthetic unity": it is a representation of a unity achieved through synthesis.

Furthermore, recall that combination (as opposed to mere synthesis) is an act that represents the "synthetic unity" that it brings about, and that this was the feature of combination that thus far has lacked an explanation. Kant's identification of pure apperception with a representation of synthetic unity suggests, then, that pure apperception is meant to fill this explanatory gap. We should therefore look for something in Kant's account of pure apperception that explains *how* it can play this explanatory role—that can explain how combination is *both* a relating together of diverse representations and a representation of the unity it thereby achieves. I believe we find this, although it is not developed as much as one might have hoped or expected. On my view, what is distinctive about pure apperception—and what enables it to fulfill this explanatory role—is that it is a consciousness of our synthesizing activity that is *constitutive of the very activity of which it is a consciousness.*

The textual basis for this interpretation begins with this passage:

> [T]his thoroughgoing identity of the apperception of a manifold given in intuition contains a synthesis of the representations, and is possible only through the consciousness of this synthesis. … The latter relation…does not come about by my accompanying each representation with consciousness, but rather by my **adding** one representation to the other and being conscious of their synthesis. (B133; original boldface, my underline)

What I take Kant to be saying is that certain acts of synthesis—the acts I'm calling 'combination'—are attended by the subject's awareness of the activity of synthesis itself, of the *doing*, which, by virtue of relating together and thereby uniting a manifold of diverse representations, constitutes a unified whole. As a first approximation, we can say that pure apperception is a mode of awareness whose object is not any particular content, but rather the subject's relating that content together in a way that constitutes a unified whole.

However, calling the activity of synthesis the "object" of our awareness invites a serious misunderstanding. The subject's relation to her synthesizing activity cannot take the form of higher-order "object" awareness. At the risk of belaboring the point, this is because such an account could not explain how we represent the unity of the relevant object, but only presuppose that we can. For this reason, the subject's awareness of the activity of synthesis must be constitutive of that activity. Relatedly, this awareness is a form of *agential* consciousness. In pure apperception I am not, and I do not represent myself as, a spectator to synthesis. Rather, through pure apperception I represent the activity of synthesis in a way that constitutively identifies me as the agent of the activity and thus constitutively identifies my activity as *my* activity.

The account is perhaps best illustrated in view of the explanatory challenges of the preceding sections. Consider first the possibility of self-consciousness. I said that the self-consciousness expressed by "I think" is a representation of unity insofar as it represents some manifold of representations as together "mine." We can now say that to represent them as mine is to be conscious of them as the elements of a synthesis of which *I* am the agent. For example, in judging that *the sun caused the stone to be red and hot* I relate together the various concepts that constitute this judgment. This representational act is a synthesis that constitutively includes an awareness of myself as its agent. Crucially, such awareness involves my identifying the constituent representations—the concepts [sun], [stone], and so forth—as the elements that I actively "hold together" in the judgment. This example illustrates a formal feature common to every combination: for every constituent representation, I represent myself as the agent responsible for determining the relations that hold between it and other constituent representations. That is what it is to represent myself as the subject of the representation, or what amounts to the same, to be conscious of the representation as "mine."

Consider next the subject's ability to represent the unity of an object. To recap, representing an objective unity is relating representations together in a way that (i) purports to have objective

significance and (ii) grounds a possible analysis. But analyzing a representation requires an awareness of how its constituent representations are related, and partly on this basis we determined that combination constitutively includes a representation of the unity of the manifold so combined. What was missing from our previous account was any awareness on the part of the subject concerning *how* the synthetic act constituted the unity of the manifold. The account of pure apperception as a form of agential consciousness fills this gap. To represent the unity of a manifold of representations is to relate together its representations, and in so doing, be *conscious of the act of relating*. In other words, my representation of the unity of a manifold of representations is made possible by my awareness that *I relate these representations to each other* in a certain way. Once more, this consciousness is internal to the synthetic act.

The provision that I am aware of relating them "in a certain way" is crucial. It highlights the fact that pure apperception is an awareness of *how* the act of synthesis achieves the unity that it does. This feature of pure apperception is made explicit in the A edition of the Transcendental Deduction.

> [T]his unity of consciousness would be impossible if in the cognition of the manifold the mind could not become conscious of the identity of the function by means of which this manifold is synthetically combined into one cognition. … [T]he mind could not possibly think of the identity of itself in the manifoldness of its representations…if it did not have before its eyes the identity of its action… (A108, my emphasis)

We see that, according to Kant, pure apperception includes the subject's awareness of the "identity of the function" of the synthesis she performs. By "function" I believe Kant has in mind how the act is structured to achieve a particular kind of unity.[33] The implied contrast is with an activity of synthesis that does not include the subject's consciousness of the nature or identity of her synthetic activity. For example, in judging that *the stone is red*, I actively relate the elements together in the form of a categorical judgment ("S is p"). The form of categoricity is one of twelve logical forms that Kant lists in his table

[33] Cf. Longuenesse (*Kant and the Capacity to Judge*, p. 3 fn. 2).

of judgments, each of which is associated with an *a priori* function of the understanding.[34] In the case at hand, my awareness of myself as the agent of synthesis includes my awareness of the structure of my act—in this case, that it is a relating of a subject with a predicate. Moreover, the concepts that occupy the subject and predicate position ([stone] and [red] in our example) are themselves associated with a rule that specifies their application conditions in terms of the sensible features of intuitions.[35] My representation of how the elements of the judgment are related is thus a representation of how *I* relate them together. Without a representation of this sort, I could not take my judgment *the stone is red* to signify an objective relation of substance and accident. That is to say: (1) it is through my awareness of being the agent of a synthesis according to a certain function that I represent the unity of my judgment *the stone is red*; (2) my ability to represent the unity of my judgment *the stone is red* is a necessary condition of my ability to represent the unity of the object represented by the judgment (i.e., the unity of the object insofar as it a red stone); and (3) my awareness of being the agent of a synthesis according to a certain function is what is expressed by "I think" in the possible judgment *I think the stone is red*.

In light of the foregoing account, a question arises as to exactly which activities of synthesis are constitutively attended by pure apperception, that is, which forms of synthesis are *combinations* in the above sense. The fact that "I think" is the generic representation of pure apperception suggests that *all activities of combination are activities of thinking*. In other words, all combination consists in the representation of a unity expressed by the relation of concepts in acts of judgment.[36] Further support for this claim is found in the B Paralogisms chapter, where Kant asserts that "I is *only* the consciousness

[34] Cf. KrV A70/B95.

[35] Cf. A106, A126, A320/B377. See also footnotes 20 and 23.

[36] I take this to include exercises of the capacity for *inference*, as a necessary concomitant of our capacity for judgment. There is a straightforward sense in which syllogistic reasoning is a representation of unity: it is the representation of the unity of distinct judgments according to a particular syllogistic form. For reasons of space I do not elaborate on the relationship of inference to judgment. Cf. A303/B359-A305/B361.

of my thinking" (B413, my emphasis). Since, according to my view, the consciousness of oneself as subject, as "I," is the pure apperception that constitutively attends every act of combination, the claim that "I" expresses *only* the consciousness of my thinking entails that every act of combination is an act of thinking. Conversely, inasmuch as every judgment is a representation of unity, every judgment is attended by the pure apperception of the activity of judging, and thereby of the "function" of this activity in achieving a certain unity among one's representations. This entails that *all activities of thinking are activities of combination*. The conjunction of these claims entails that thinking and combining are one and the same activity.

I anticipate two objections to this last claim, which pull in somewhat opposite directions. According to the first objection, if all combination is thinking, then I owe an explanation of how subjects can represent the unity achieved through a "mere" synthesis—for example, the spatiotemporal ordering of sensation in an intuition that was discussed in §1.3. Kant asserts that we "apprehend" the manifold of an intuition "as manifold," that is, in such a way that we discriminate the spatiotemporal ordering of its matter (sensation) and thereby represent its spatiotemporal unity (A98-100).[37] But if the unity of an intuition is achieved through a prediscursive synthesis of which we are not conscious, and if the consciousness of the activity of synthesis is precisely the feature of combination that makes it a representation *of unity*, then—so the objection goes—it becomes a mystery how we are able to represent the unity of that intuition. A full explanation of our capacity to represent the unity of our intuitions would take us beyond the scope of this chapter, but in response I can say the following. What is entailed by the identification of combination with thinking is not that we *never* represent the unity of our intuitions, but rather that *in* representing the unity of our intuitions we are conscious of the conceptual forms of their synthesis, and therefore, according to my account, of ourselves unifying the manifold of sensibility according to those conceptual forms. Those conceptual

[37] This is in addition to an *a priori* synthesis of space and time as pure forms of intuitions. Cf. A99-100.

forms correspond, of course, to the categories of the understanding, as the concepts that give the manifold of sensibility its "necessary synthetic unity."[38]

The second objection concerns the fact that, on my interpretation, thinking is necessarily self-conscious: in claiming that all thinking is constitutively attended by the pure apperception of its activity, I do not claim merely that the activity of thinking *can be* made self-conscious, but rather that it is *always* so. But this could be seen as requiring more than the principle of the analytical unity of apperception demands. That principle states merely that the representation "I think" *can* accompany all my representations insofar as those representations are "something for me," their subject. Therefore, the objection goes, it is natural to suppose that on Kant's view, thinking *can be* made self-conscious in an act of pure apperception, not that it necessarily is so. There are two points to make in response. The first is that, in allowing for the possibility of a "prediscursive" synthesis of our intuitions of which we are not conscious, my account is compatible with the existence of cognitively significant representations that are not apperceived. When Kant introduces the principle of the analytical unity of apperception in §16 of the B Deduction, he emphasizes *intuitions* as the sort of representations that merely "can" be accompanied by the representation "I think": "all manifold of intuition has a necessary relation to the I think in the same subject in which this manifold is to be encountered" (B132). My account provides a way of interpreting the "necessary relation" that intuition bears to the representation "I think": that relation does not consist in every intuition *actually being apperceived*, but rather in the *possibility* of every intuition being apperceived in thought, through the process of

[38] This observation points toward a way of understanding Kant's claim, at B131, that in combination we "add" a representation of unity to the representation of the manifold. According to this interpretation, the representation of unity that is "added" to the manifold is a categorial form that "guides," so to speak, a particular representation of unity, by subsuming the manifold under that category. For example, a judgment that *snow is white* is made possible by thinking of the manifold in terms of subject and predicate relations. On this view, we should regard the judgment *snow is white* as the particular resultant unity guided by the form of categoricity ("S is p"), which thereby serves as an orienting principle of synthesis that is itself a representation of unity. Such an interpretation is defended by Longuenesse (*Kant and the Capacity to Judge*) and Boyle ("Kant's Categories as Concepts of Reflection" (unpublished)). Here I would like to express my gratitude to both of them for helping me see the connections of their views to my account of pure apperception.

conceptualization of the sensible manifold described in the preceding paragraph. The second point is that while the representation "I think" is a discursive representation of pure apperception, it is *not* a necessary concomitant of pure apperception. So while we are, in thinking, always aware of the activity of thinking, specifically of ourselves as the agents of such thinking, this does not entail that every act of thought is explicitly accompanied by the representation "I think." Indeed, I take it that Kant believes that in the vast majority of cases, self-conscious subjects are attending not to their mental agency but rather to the content of what is thought. Kant allows, moreover, that self-consciousness comes in degrees (B414-415). Thus, the claim that all thinking is constitutively attended by pure apperception, and that such self-consciousness is what makes thinking a representation of unity, is not an implausibly restrictive requirement on our capacity for thought, for it is ultimately compatible with pure apperception being a merely implicit feature of much of our thinking.

I will close this chapter by bringing to the foreground a crucial implication of my account of pure apperception. I have argued that pure apperception is the subject's consciousness of being engaged in the activity of combination, which turned out to be identical to judgment,[39] the activity of thinking. We saw that within Kant's taxonomy of representations and their relation to our cognitive capacities, judgment belongs to the understanding, "the faculty for bringing forth representations itself, or the spontaneity of cognition" (A51/B75). That is why Kant claims, in his discussion of the principle of the analytical unity of apperception, that the representation "I think" is an expression of the subject's spontaneity (B132). Likewise, in a section of Kant's *Anthropology from a Pragmatic Point of View* (1798) titled "On inner sense," Kant characterizes the difference between pure apperception and inner sense as the difference between consciousness of doing something [*etwas tun*] and consciousness of undergoing something [*etwas leiden*] (An 7:161). Pure apperception is therefore our consciousness of being engaged in exercising our rational spontaneity. But since, in pure apperception, we are, in

[39] Here again I mean to include inference as falling within our capacity for judgment. See footnote 36.

addition to being conscious of ourselves as the agents of a spontaneous synthesis, conscious also of the *function* of that synthesis, we are at least implicitly aware of *what it is* we're doing when we exercise our rational spontaneity, and this is true across different exercises of rational spontaneity.[40] We can think of our various capacities for thought as each individuated by an internal principle, one that supplies the constitutive standard or "rule" for that capacity. To cite an earlier example, one sub-capacity within our overall capacity for thought is our capacity for categorical judgment, which is internally guided by the rule for ordering concepts in subject-predicate form.[41] Such a rule supplies the standard for *what it is* to form a categorical judgment; to exercise a capacity that is not guided by that rule is *thereby* not to exercise the capacity for categorical judgment.[42] For this reason, Kant's claim that in pure apperception we are conscious of the "function" of some synthesis can be described as follows: whenever we are thinking, we are at least implicitly aware of the internal principles that govern and individuate our various capacities for thought, and of our own status as the agent of those capacities. This in turn suggests that by focusing our attention on those principles, we can achieve a more complete and systematic cognition of our capacity for thought, including our capacity for the discursive cognition of objects. I take up this suggestion in the next chapter.

---

[40] Here I have in mind consciousness of the function of (i) different particular judgments (e.g., judging that *snow is white* versus judging that *snow is cold*), (ii) different kinds of judgments (e.g., categorical judgment versus hypothetical judgment), (iii) different kinds of exercises of spontaneity within our overall capacity for reason (e.g., judgment versus inference), and so forth.

[41] See footnote 38. In referring to a "sub-capacity" I mean to imply that capacities can be understood as nested within one another: a particular capacity might resolve into sub-capacities while itself serving as a sub-capacity for some even more general capacity.

[42] The role of constitutive standards in Kant's philosophy has been highlighted by Christine Korsgaard, who defines them as "standards that apply to a thing simply in virtue of its being the kind of thing that it is" (*Self-Constitution: Agency, Identity, and Integrity*, p. 28). This idea has been echoed by Reath: "We may think of the *form* of some rational activity or object of cognition as the constitutive or defining features of an activity of entity of that kind—the features that an activity or entity must possess to count as an instance of that kind. The form of some rational activity or object of cognition will be associated with a *formal principle* that is constitutive of that rational activity or object of cognition. The formal principle of some rational activity would be the guiding internal or constitutive norm that a subject must follow in order to engage in that activity. By specifying the form of that activity, it provides a norm that anyone engaged in that activity must satisfy and that in some sense does guide any instance of the activity" ("Formal Principles and the Form of a Law," p. 42).

CHAPTER TWO

Reflection and *Critique* in Kant's Theoretical Philosophy

§2.0 The previous chapter provided an account of Kant's conception of self-consciousness. According to that account, there is a form of self-consciousness (or "apperception") that cannot be explained on an empirical-observational model. This "pure" mode of apperception consists in the subject's awareness of herself as the agent of the activity Kant calls 'combination', which was identified with thinking, the act of relating concepts together to form judgments. I said that such self-awareness is a constitutive feature of this activity. Unlike empirical apperception, pure apperception does not consist of a second-order representation, one that takes the subject's own mental occurrences as its objects. That is, pure apperception is not a form of self-awareness in which the self appears to the subject in the form of some given representational content. Rather, thinking is an activity that constitutively includes the subject's awareness of herself as the "I" that does the thinking, the "agent" of thought.

I argued that this account of self-consciousness performs double duty for Kant. First, it enables Kant to explain the basis for the representation "I think" compatibly with the observation that no given content could possibly serve as the object of this representation. Second, it enables Kant to explain how we can represent the unity of an object in such a way that that very representation can serve as the basis for an *analysis* of that object. Recall that a representation (e.g., the judgment *this stone is hot*) can serve as the basis for an analysis of its object through its constitutively including an awareness of how its constituent representations (e.g., the concepts [stone] and [hot]) are related. I said that this feature is explained by the fact that pure apperception includes the subject's (perhaps implicit) awareness of the rule or "function" of combination, the way in which discrete representations are combined by the subject to form a unified whole.

We might nonetheless wonder what justification Kant has for claiming, apart from the role of pure apperception in the above explanations, that we are capable of such self-consciousness in the first place. In particular, what entitles us to the belief that we are the kinds of subjects to whom the principle of the analytical unity of apperception applies? If we assume that we possess a non-empirical mode of self-consciousness, then there will be a real question about what this capacity amounts to, to which the account of pure apperception provided in the previous chapter gives an attractive answer. But this assumption will strike the modern reader, as it would have struck Kant's empiricist opponents, as controversial in its own right. Questions therefore arise about the ultimate justification of this doctrine, as well about Kant's philosophical methods more generally.

But given that Kant's doctrine of pure apperception is a theory concerning the subjective nature of *our* self-consciousness, the very demand to prove, through philosophical argumentation, its existence in us, might seem incongruous with its topic. First, it's not clear what kind of discursive "proof" could be offered for a claim of this sort. Second, such a proof would hardly be needed. For if Kant's doctrine of pure apperception were true, then the fact that we possess a capacity for pure apperception would be available to us as an item of self-cognition just in virtue of the fact that we are thinkers at all. The pure apperceptive nature of thought entails, for example, that in counting to ten or subsuming some given intuition under an empirical concept, I am in each case at least implicitly aware of *my own engagement* in a synthesis of concepts according to a rule. I could, by deliberately attending in pure apperception to my role as an active thinker, thereby make the pure apperceptive character of my thought an explicit item of self-cognition. In other words, the *doctrine* of pure apperception could be justified by reflectively attending to my own thinking activity in precisely the manner that the *fact* of pure apperception makes possible.

My aim in this chapter is to relate Kant's account of pure apperception to the above method, which going forward I will call 'reflection'. In §2.1, I will present a brief outline of Kant's conception

of critique, arguing that critique is neither a form of conceptual analysis nor a form of empirical cognition. In §2.2, I will argue that, on Kant's view, reflection is a source of self-cognition that is indispensable to the task of critique, and that reflection is itself a mode of pure apperception. Moreover, Kant regards the doctrine of pure apperception presented in the previous chapter as a fundamental item of reflective self-cognition. Thus, while reflection is a crucial component of the method through which we are able ultimately to justify and articulate the doctrine of pure apperception, our capacity for reflection is itself grounded in the pure apperceptive structure of thought. There is therefore an internal connection between the *method* of Kant's critique of theoretical reason and a key principle that this critical procedure articulates and defends. Finally, in §2.3, I will argue that the task of reflection is not contingently related to the concept of a critique of theoretical reason, but rather a necessary aspect of any such critique insofar as critique itself presupposes and exhibits the autonomy of reason. The material from this last section will become especially relevant in chapter four, where I will argue that the doctrine of the fact of reason indicates Kant's reliance on a *practical* mode of reflection, and that Kant's presumption of a capacity for practical reflection likewise expresses Kant's commitment to the autonomy of reason.

§2.1 I mentioned in the previous chapter that one of the chief tasks of the *Critique of Pure Reason* is to establish the possibility of synthetic, *a priori* cognition. I would now like to orient this project with respect to the project of a "critique" of theoretical reason and elaborate on the methods and presuppositions of this project. This will set the stage for introducing *reflection* as an essential component of critique.

At the outset of the A-edition Preface to the first *Critique*, Kant makes the following pronouncement: "Human reason has the peculiar fate in one species of its cognitions that it is burdened with questions which it cannot dismiss, since they are given to it as problems by the nature

of reason itself, but which it also cannot answer, since they transcend every capacity of human reason" (Avii). The questions that cannot be dismissed, and which issue from reason itself, are the traditional questions of metaphysics, such as whether there is a God or whether human beings possess freedom of the will. Kant makes this pronouncement against a backdrop of doubt on the part of certain of his contemporaries about the validity of metaphysics as a science. Such doubt, he claims, is warranted by the fact that metaphysics has not yet established for itself systematic and justified principles and procedures for addressing its questions, and moreover has a tendency to produce contradictory answers. A major task of the first *Critique* is to finally validate—while at the same time identifying the limits of—metaphysics as a theoretical enterprise. Because Kant conceives of metaphysics as a science of *pure* reason, Kant's interest in the validity of metaphysics as a well-grounded science gives rise to the more general question that animates much of the first *Critique*: How is synthetic, *a priori* cognition possible?[1] In answering this question, Kant aims to secure a philosophical foundation for (limited, it will turn out) metaphysical inquiry.

For Kant, the question of the possibility of synthetic, *a priori* cognition concerns the scope and legitimacy of pure reason in its theoretical exercise. Kant insists that metaphysics cannot be established as a science without an antecedent examination of the principles of pure theoretical reason, the source of their legitimacy, and the scope of their legitimate application. Such an examination is the project of a *critique* of pure theoretical reason.

Kant's critical-period works are suffused with references to critique, but he presents what is perhaps his most comprehensive statement concerning the project of critique in the A Preface, shortly after making the proclamation cited above:

> [The] power of judgment [of our age], which will no longer be put off by illusory knowledge…demands that reason should take on anew the most difficult of all its

[1] It is a more general question at least in the sense that it is relevant not only to metaphysics but to all branches of human inquiry that presuppose or issue in synthetic, *a priori* judgments, which for Kant include mathematics and natural science.

> tasks, namely that of self-knowledge [*Selbsterkenntnis*],[2] and to institute a court of justice, by which reason may secure its rightful claims while dismissing all its groundless pretensions, and this not by mere decrees but according to its own eternal and unchangeable laws; and this court is none other than the **critique of pure reason** itself.
>
> Yet by this I do not understand a critique of books and systems, but a critique of the faculty of reason [*das Vernunftvermögen*] in general, in respect of all cognitions after which reason might strive **independently of all experience**, and hence the decision about the possibility or impossibility of a metaphysics in general, and the determination of its sources, as well as its extent and boundaries, all, however, from principles. (Axi-xii)

This passage identifies three important features of critique. They are:

(1) Critique as reason's *self*-examination. The term 'critique' appears throughout Kant's critical writings, with Kant emphasizing different aspects of critique depending on his purposes at the time.[3] However, the project of critique is most essentially the philosophical examination of reason, for the purpose of legitimating some (possibly disputed) claim about reason's powers of cognition or identifying the boundaries of reason's legitimate use. Since it is through the use of reason that we engage in the project of critique, its task is inherently reflexive, bringing reason to bear upon itself. A successful critique therefore culminates in a kind of self-cognition [*Selbsterkenntnis*], which Kant identifies with the aim of critique in the passage quoted above (Axi). In particular, critique issues in the subject achieving cognition about the validity and scope of her cognitive powers.[4] The notion that

---

[2] As I explain below, I will use the term 'self-cognition' in place of 'self-knowledge' going forward.

[3] In particular, critique can serve both a positive and negative function, in that it can both (i) establish the possibility of a particular kind of cognition and (ii) set limits on the scope of that cognition. Kant will emphasize one or the other of these functions in different contexts. The Preface to the *Groundwork of the Metaphysics of Morals* emphasizes the positive aspect of critique. There the task of a "critique of pure practical reason" is identified with that of securing a foundation for a "metaphysics of morals," a system of moral first principles and duties. The *Critique of Practical Reason* begins with a methodological remark that invokes the negative aspect of critique: pure practical reason doesn't require a special critique, according to Kant, because there is no worry that it might venture beyond its legitimate use (KpV 5:3). For helpful commentary on the notion of "critique" and its role within Kant's philosophy, see Dieter Henrich ("The Deduction of the Moral Law: The Reasons for the Obscurity of the Final Section of Kant's *Groundwork of the Metaphysics of Morals*," pp. 308-311; and "Kant's Notion of a Deduction and the Methodological Background of the First *Critique*") and William Bristow (*Hegel and the Transformation of Philosophical Critique*, pp. 53-61).

[4] Strictly speaking, Kant claims that *reason* achieves self-cognition through critique. I take this claim to be a somewhat grandiose formulation of the thought that critique is *an exercise of reason* through which the rational subject achieves cognition *of her own reason*.

critique is an essentially reflexive procedure is hinted at elsewhere in Kant's corpus. For example, in the *Groundwork of the Metaphysics of Morals* Kant claims that establishing the authority of morality would require going "beyond cognition of objects to a critique *of the subject*" (G 4:440, my emphasis).

(Here I must offer a brief comment about terminology. In the passage cited above, which was taken from the Cambridge Edition of the first *Critique*, '*Selbsterkenntnis*' is translated as 'self-knowledge'. But in keeping with the practice of translating '*Erkenntnis*' as 'cognition', I will adopt the term 'self-cognition' going forward. This serves to bring the discussion into alignment with other translations of Kant's texts. For example, in the Cambridge Edition of the *Jäsche Logic*, the study of logic is characterized as "a self-cognition [*Selbsterkenntnis*] of the understanding and of reason, not as to their faculties in regard to objects, however, but merely as to form" (JL 9:14). As we will see in §2.2, the task of critique and the enterprise Kant calls 'general logic' share a common method: both are made possible through reflection on our rational powers. While a successful critique produces a kind of self-cognition, the activity of reflection itself yields an unargued-for cache of items of self-cognition on the basis of which critique can be undertaken, i.e., that serve as the needed premises of a possible critical investigation. Importantly, Kant tends to use the term 'knowledge' (*Wissen*) in connection with a notion of universal validity that is grounded in the *objects* of our cognition.[5] By contrast, the self-cognition of reflection—and therefore, by extension, of critique—is a kind of "pure apperceptive" cognition, that is, cognition of ourselves "as subject." Thus, adoption of the term 'self-cognition' in place of 'self-knowledge' may serve to remind us that reflection and critique yield a kind of cognition that is distinct from "objective" cognition in the above sense. But even with this qualification in place, care must be taken, for in the first *Critique* Kant tends to reserve the term 'cognition' precisely for representation of *objects*.[6] For this reason, unless I explicitly indicate otherwise, going forward

[5] Cf. A820/B848-A823/B851.

[6] Cf. A320/B376-377.

'cognition' will refer to objective representation and 'self-cognition' will refer to the cognition of ourselves as thinking subjects.[7])

(2) Critique as concerned with the sources and limits of our cognition. I said just now that critique has the twofold function of legitimating our powers of cognition and identifying the boundaries of their legitimate use.[8] This twofold function is reflected in the above passage's statement that critique is concerned with "the determination of [the] sources, as well as [the] extent and boundaries" of pure reason (Axii). As we will see in the next chapter, we legitimate some cognitive power in part by revealing the "source" of the relevant mode of cognition—for example, in the claim that our synthetic, *a priori* knowledge of the natural world is grounded in a synthetic unity of apperception. This is the task within critique that Kant refers to as 'deduction'. In establishing the legitimacy of some mode of cognition, critique can then serve as the foundation for a further systematic elaboration of pure reason's principles.[9] However, in identifying the source of some cognitive capacity, we may also discover that it is restricted in its application. No doubt the most

[7] This is almost entirely a matter of terminological stipulation in an attempt to align our terms with Kant's own somewhat inconsistent phraseology. It would not be wrong to think of Kantian *Selbsterkenntnis* as "self-knowledge" in the contemporary sense of that term, again provided that one keeps in mind that it is not a form of "object knowledge." It has been suggested to me that 'self-cognition' might be the more appropriate term even apart from Kant's phraseology, because 'knowledge' suggests a justificatory relation. However, while reflection on our cognitive capacities does not provide any sort of justification that is grounded in the *objects* of our knowledge, that does not entail that reflection is not itself the ground of a genuine justification to hold certain facts about the nature of one's own subjectivity to be true for the purposes of critique. In the case of objective cognition (i.e., objective representation), Kant seems to allow for a distinction between cognition and knowledge. Specifically, he refers to the fact that we can have false cognition (A58/B83), but he presumably does not think we can have false knowledge. By contrast, I have found no evidence that Kant regards the "self-cognition" particular to critique to be the sort of thing that could be *false* in the way that our representations of given objects can be. Indeed, Kant suggests at one point that the nature of our understanding (the topic of critical investigation in the Analytic of the first *Critique*) "cannot be hidden from us," which seems to presume the transparency of certain aspects of the mind to itself (A12-13/B26).

[8] See footnote 3.

[9] Cf. A11/B25-A16/B30.

notorious "restriction" in Kant's philosophy is the limitation of theoretical knowledge to appearances.[10]

(3) Critique as presupposing and exhibiting the autonomy of reason. Kant conceives of the critique of pure reason by analogy with a "court of justice" instituted by reason itself, where the laws of this court are "[reason's] own eternal and unchangeable laws" (Axi-xii). If we take this metaphor seriously, we see that critique therefore exemplifies in crucial aspects the conception of rational autonomy that Kant lays out in his practical philosophy. There the categorical imperative is presented as both the highest principle of morality and the internal principle of the agent's *own* pure practical reason. For example, the *Groundwork* tells us that "the human being is bound to laws by his duty, but…*subject only to laws given by himself but still universal*" (G 4:432). This is an instance of "the principle of the **autonomy** of the will," the general precept that "the will is in all its actions a law to itself" (4:433; 4:447).[11] To be sure, Kant does not claim (and does not hold) that critique is an exercise of *will*, or practical reason.[12] Nonetheless, the relevant point for our purposes is that autonomy consists in the agent's guiding her activities not by external edicts but by the "laws," or principles, of her own reason. Critique therefore appears to be a philosophical method that Kant conceives on this model. In its

[10] Similarly, the *Critique of Practical Reason* offers a critique of empirical practical reason in the negative sense of restricting the boundaries of its legitimate deployment. Kant argues that empirical practical reason must meet a condition of universalizability, the standard of which is provided by the Formula of Universal Law.

[11] Cf. KpV 5:33ff.

[12] Kant identifies the will with practical reason at G 4:412. In "Reason and Politics in the Kantian Enterprise," Onora O'Neill argues that the categorical imperative (understood in terms of its first formulation, as the imperative only to adopt maxims whose universal adoption one could concomitantly legislate) is the highest principle of reason, and thus is the highest principle of reason's *critique*. In this way, critique is conceived as a *practical* task, one governed by pure practical reason's most fundamental principle (pp. 3, 18ff.). Here I will outline my main reasons for rejecting O'Neill's conclusion. First, as I argue in chapter four, Kant understands practical reason in terms of its formal relationship to its object: practical reason is *productive* in regard to its object. By contrast, while the decision *whether* to undertake a critique of reason is a practical one, critique itself doesn't produce an object on the basis of its representation of one. But that does not entail that critique is a *theoretical* task: as I argue in chapter four, critique is neither practical nor theoretical in Kant's narrow senses of those terms. Furthermore, O'Neill claims that critique does not ultimately proceed from any predetermined method—for example, critique is not based on a Cartesian method of introspection—but rather requires a kind of "coordination" of rational subjects in the methods they adopt, methods that are always, in principle, revisable. Against this, I claim that critique is made possible through first-person reflection on our own cognitive capacities, where reflection is to be understood *not* as a process of introspection but rather as a mode of pure apperception.

fundamental methodological orientation, critique institutes an Enlightenment standard according to which rational subjects are normatively constrained only by the laws of their own reason, where no putative authority is exempt from rational assessment.[13] Reason's self-critique therefore does not consist in its recognition that principles of pure reason either meet or fail to meet some external standard of assessment. Rather, questions about the scope and legitimacy of pure reason are "resolved…to reason's full satisfaction," according to rationally self-prescribed standards (KrV Axiii). In this way, critique operates from a presupposition of the absolute normative independence of reason, and exhibits this independence through its operation.

So far I have characterized critique only with regard to its defined general features. How exactly does it work? We must suppose that the rational principles operative in critique would include the principles of formal logic, in particular, the law of non-contradiction: whatever else might be said of critique, its conclusions must be logically consistent. This conception of critique coheres nicely with the idea that critique is an activity guided by reason's own internal principles and standards of assessment. For, in Kant's time, the subject matter of logic was not understood as it is by contemporary logicians, that is, in terms of the formal properties of propositions, relations of truth-preservation that hold between propositions, and the rules constitutive of the derivation systems we adopt. Instead, logic was regarded as the study of the rational operations of the mind.[14] Consider, for example, Kant's definition of what he called 'pure general logic': "A **general** but **pure** logic…has to do with strictly *a priori* principles, and is a **canon of the understanding** and reason, but only in regard to what is formal in their use, be the content what it may (empirical or transcendental)" (A53/B77). Within Kant's taxonomy of logic—about which I will have more to say later—pure general logic is in

[13] This is seen in a footnote that immediately precedes this discussion: "Our age is the genuine age of criticism, to which everything must submit. **Religion** through its **holiness** and **legislation** through its **majesty** commonly seek to exempt themselves from it. But in this way they excite a just suspicion against themselves, and cannot lay claim to that unfeigned respect that reason grants only to that which has been able to withstand its free and public examination" (KrV Axi).

[14] Cf. Longuenesse (*Kant and the Capacity to Judge*, pp. 5, 74).

a certain sense the closest analog to the contemporary understanding of logic, for it concerns the formal principles of the understanding in abstraction from any consideration of how our thought relates to an object, how our thought has *content*. The idea is that, for example, in the universal, affirmative, categorical judgment "all Xs are Ys," the understanding is internally guided by the rule that excludes the particular, negative, categorical judgment "some Xs are not Ys." This rule is a principle of general logic because it "abstracts from all contents of the cognition of the understanding and of the difference of its objects, and has to do with nothing but the mere form of thinking" (A54/B78). Thus, on the proposal under consideration, the principles of critique include the principles of pure general logic, conceived as formal constitutive principles of our capacity for thought.

While it is no doubt true that the principles of pure general logic are deployed in critique, critique itself must include sources of (self-)cognition beyond what pure general logic can provide. Principles of pure general logic can at most logically order and systematize some set of true propositions, and tease out their formal logical presuppositions and implications. Together with knowledge of the content of our concepts, pure general logic can guide and regulate conceptual analysis. But critique aims to establish synthetic, *a priori* truths, and no exercise in mere conceptual analysis can justify a synthetic, *a priori* truth except on the basis of some other synthetic, *a priori* truth. Likewise, experience cannot be the source of such cognition, for experience yields cognition only of synthetic, *a posteriori* truths. And while certain commentators have interpreted Kantian critique as an exercise in conceptual analysis or as relying on experiential cognition,[15] a passage from Kant's *Groundwork* explicitly tells against such interpretations. The context for this passage is that Kant has just argued that the categorical imperative is a principle of autonomy, but he has not yet shown that any actual rational agent is obligated to follow it. Kant informs us that since the latter claim is a

[15] See the discussion in Smit ("The Role of Reflection in Kant's *Critique of Pure Reason*," pp. 204-205).

synthetic, *a priori* proposition, it cannot be established on the basis of either conceptual analysis or "cognition of objects," by which he means experience:

> That this practical rule is an imperative, that is, that the will of every rational being is necessarily bound to it as a condition cannot be proved by mere analysis of the concepts to be found in it, because it is a synthetic proposition; one would have to go beyond cognition of objects to a critique of the subject, that is, of pure practical reason, since this synthetic proposition, which commands apodictically, must be capable of being cognized completely a priori. (G 4:440, my emphasis)

Here we see that Kant conceives of critique as a source of justification that, precisely because it issues in synthetic, *a priori* truths, must be distinguished from both conceptual analysis and experiential cognition. And while this passage is found in Kant's practical philosophy, there is nothing distinctively "practical" undergirding the point he makes, which appears to apply to critique in general and not just the immediate task of a critique of pure practical reason.

All of this implies that reliance on a non-experiential source of self-cognition is indispensable to the task of critique, for without such a source there would be nothing to distinguish critique from conceptual analysis or empirical cognition. And since critique is a task of the self-cognition of reason, it is natural to suppose that critique draws upon, and therefore methodologically presupposes, a non-experiential source of self-cognition on the basis of which the rational subject can achieve *further* self-cognition, in particular, cognition of the scope and legitimacy of her cognitive powers. I take up this proposal in the next section, where I discuss Kant's notion of *reflection*.

§2.2 A number of commentators have in recent years converged on an interpretation according to which Kantian critique is informed by the subject's first-person *reflection* on her cognitive capacities.[16] Although these commentators differ in crucial respects, and invoke Kant's notion of *reflection* to serve

[16] Such commentators include Henrich ("Kant's Notion of a Deduction and the Methodological Background of the First *Critique*"), Smit ("The Role of Reflection in Kant's *Critique of Pure Reason*"), Merritt (*Drawing from the Sources of Reason: Reflective Self-Knowledge in Kant's First Critique*), and Boyle ("Kant's Categories as Concepts of Reflection" (unpublished)).

different interpretative aims, the basic idea that unites their accounts is the following: implicit in ordinary, first-order cognition is the pure apperceptive consciousness that, when explicitly drawn into focus, can provide the subject with the basis of self-cognition that I have claimed is indispensable to the task of critique.

This section has three main goals. Its first goal is to elucidate and expand upon what is meant by the claim that the critique of theoretical reason has its basis in reflection on our rational powers. To that end, I will make extensive use of interpretations provided by Dieter Henrich and Houston Smit, whose accounts of the first *Critique*'s methodological reliance on reflection I endorse. (Questions about or disagreements with their accounts will be acknowledged in footnotes. They are incidental to the aims of this section. As we will see, Smit's account is in large measure an expansion of Henrich's.) The second goal of this section is to reinforce their accounts by providing textual evidence that the relevant notion of reflection in fact has its basis in our capacity for pure apperception. Finally, the third goal of this section is to highlight the relationship of the *method* of reflection to the first *Critique*'s *doctrine* of pure apperception. I will argue that Kant expects us to assent to the first *Critique*'s doctrine of pure apperception on the basis of our reflecting on our own powers of cognition. In other words, while our capacity for reflection is grounded in our capacity for pure apperception, the *doctrine* of pure apperception is an item of self-cognition made available through our reflecting on our own case. This will set the stage for §2.3, where I will argue that Kant's reliance on reflection is ultimately rooted in his conception of critique as an exercise of our rational autonomy.

Before proceeding, it bears mentioning that Kant uses the term 'reflection' (*Überlegung*) in distinct ways across different contexts, and there is some dispute among commentators concerning what connections, if any, these uses bear to one another.[17] First, Kant claims that "the understanding

[17] Beatrice Longuenesse has claimed that reflection conceived as a power operative in concept generation (i.e., a capacity for "universalizing" a feature of, e.g., a sensible manifold by way of introducing a concept that marks that feature) is distinct from "transcendental reflection," understood as the subject's self-consciousness of her cognitive powers and the

intuits nothing, but only reflects" (Prol 4:288). Since Kant elsewhere defines the understanding as the capacity for thought—that is, the capacity for deploying concepts in acts of judgment (KrV A69/B94)—it seems to follow that Kant identifies some notion of reflection with thinking in general.[18] In the *Jäsche Logic*, Kant defines reflection as a special power of the understanding through which it (along with the powers he calls 'comparison' and 'abstraction') generates concepts (JL 9:94-95).[19] In the third *Critique,* Kant makes a distinction between the "determining" and mere "reflecting" powers of judgment (KU 20:211-216). Finally, in the chapter of the first *Critique* titled "On the Amphiboly of Concepts of Reflection," Kant introduces a notion of reflection in terms of a kind of awareness of our own cognitive activities (KrV A260/B316-A263/B319).

My primary focus is on the last of these uses of the term. At the outset of the Amphiboly chapter Kant states that reflection does not concern the objects of cognition but rather the "subjective conditions" of our cognition of objects:

> Reflection (*reflexio*) does not have to do with objects themselves, in order to acquire concepts directly from them, but is rather the state of mind in which we first prepare ourselves to find out the subjective conditions under which we can arrive at concepts. It is the consciousness of the relation of given representations to our various sources of cognition, through which alone their relation among themselves can be correctly determined. (A260/B316)

He adds:

> [A]ll judgments…require a reflection, i.e., a distinction of the cognitive power to which the given concepts belong. (A261/B317)

relation of given representations to them (*Kant and the Capacity to Judge*, pp. 113-114 fn. 22). However, in unpublished work Matt Boyle has argued that these powers of reflection are aspects of a single cognitive activity ("Kant's Categories as Concepts of Reflection"). And while both of these authors regard the first sense of reflection as in part a power of generating concepts from non-conceptual materials, Melissa Merritt has argued against this interpretation in "Varieties of Reflection in Kant's Logic."

[18] Cf. Merritt ("Varieties of Reflection in Kant's Logic," p. 479).

[19] See footnote 17.

Broadly speaking, then, Kant appears to associate reflection with an awareness of our own powers of cognition and the relation of particular representations to those powers. He claims, moreover, that such awareness is a necessary condition of judgment and therefore discursive thought itself. The assertion that a kind of self-consciousness is a prerequisite of discursive cognition is both striking and familiar, for as we saw in the previous chapter, this is a key conclusion of the Transcendental Deduction of the Categories.

Before proceeding to Henrich and Smit, I must highlight certain ambiguities in Kant's characterizations of reflection, ambiguities that reappear in Henrich and Smit's own accounts. To forestall any misunderstanding, I will introduce my own regimentation of terms. This will involve in certain instances translating the positions of Henrich and Smit into my own phraseology, but not in a way that I believe alters the spirit of those positions. To begin, note that the term 'reflection' can refer either to a kind of self-cognition or, as the nominalization of the verb 'reflect', to the activity of self-consciousness through which that self-cognition is generated. (A similar ambiguity arises in the cases of 'representation', 'intuition', and 'judgment', each of which can refer either to a resultant representation (or intuition or judgment) or to the act of representing (or intuiting or judging).) We saw that Kant defines reflection in general as a "consciousness of the relation of given representations to our various sources of cognition" (KrV A260/B316). However, in the same paragraph he refers to a specifically "transcendental" mode of reflection as a kind of "action" (A260/B316-A261/B317). Kant's characterization of reflection as "a distinction of the cognitive power to which the given concepts belong" is similarly ambiguous between a resultant representation of that distinction or the act of distinguishing (A261/B317). Henrich's principal characterization of reflection is as a kind of self-knowledge.[20] By contrast, I will use the term 'reflection' exclusively to refer to the subject's activity of *reflecting* on her own cognitive capacities. I have already provided reasons for adopting the term

[20] "Kant's Notion of a Deduction and the Methodological Background of the First *Critique,*" pp. 42-43.

'self-cognition' in place of 'self-knowledge'. In light of that stipulation, and in order to capture what Henrich means by 'reflection' *qua* self-knowledge, I will instead use terms like 'self-cognition', 'reflective self-cognition', or 'item of self-cognition made available through reflection'.

Yet another ambiguity concerns reflection considered as an activity. We can think of the activity of reflection either as one that we engage in involuntarily in ordinary, first-order cognition, or as one that we deliberately undertake in order to isolate and systematize our various cognitive powers, for the purpose of critical inquiry. In the first passage cited above, Kant characterizes reflection in terms of an activity of "prepar[ing] ourselves to find out the subjective conditions under which we can arrive at concepts" (A260/B316). This phrasing suggests that Kant has in mind an activity that is undertaken deliberately, for the purpose of achieving explicit self-cognition. Likewise, I will use the term 'reflection' exclusively to refer to the activity whereby the subject *intentionally* focuses on the "subjective conditions" of her various rational activities, in order to achieve self-cognition of the sort that renders critique possible. By contrast, Henrich and Smit sometimes speak of a "natural" or "spontaneous" reflection to pick out what I have been referring to (and will continue to refer to) as the "pure apperception" that attends all thinking, which makes reflection *qua* voluntary activity possible. Thus, to sum up: (1) pure apperception (described in detail in chapter one) makes possible reflection, a mode of pure apperception in which the subject deliberately draws into focus the nature of her own thinking activity; (2) such reflection provides us with a basis of self-cognition; and (3) such self-cognition provides a basis for a possible critique of reason, reason's examination of its own sources, limits, and legitimacy.

To the best of my knowledge, Dieter Henrich was the first commentator to claim that critique draws on items of self-cognition made available through first-person reflection on one's cognitive capacities. Henrich introduces this claim by calling our attention to Kant's distinction between "reflection" and "investigation," a distinction found in the Amphiboly chapter of the first *Critique* as

well as Kant's lectures on logic.[21] Philosophical investigation (in Kant's sense) is preceded by reflection, and Henrich infers from this that reflection is a condition of the possibility of investigation.[22] So, for example, while the *Critique of Pure Reason* examines our claims to theoretical knowledge in part through an investigation of the cognitive source and legitimacy of certain *a priori* concepts and principles, that investigation is grounded in reflection on the cognitive powers themselves, and it is only against the background of such reflection that the project of critique can be undertaken.

Henrich claims that our power of reflection is made possible by the fact that our cognitive activities are attended by an awareness of the nature of the activities in which we are engaged. As stated, this account does not distinguish the activity of reflection from introspection, which might take a particular cognitive activity as its object but is distinct from the activity in question. Henrich, however, explicitly warns against such a reading: the awareness he has in mind is "not introspection," but rather "accompanies operations internally."[23] I take Henrich to mean by this that the awareness of these operations is constitutive of those very operations and, by extension, not just a precondition but moreover an essential feature of rational subjectivity. This notion—of an awareness of the nature of an activity constitutive of that very activity—specifies in part the account of pure apperception presented in the previous chapter. It was in those terms that I characterized pure apperception with respect to the activity of combination: in pure apperception we are aware of ourselves as the "agents" of combination according to a certain "function" of the understanding, that is, of ourselves combining representations to achieve a particular kind of unity among those representations. Henrich's account therefore implies an essential role for pure apperception in the method of critique: the pure

[21] Ibid., pp. 41-42.

[22] Henrich writes, "Since Kant claims that reflection precedes investigation, it is plausible to suppose that reflection is the source by means of which investigation can be undertaken" (ibid., p. 42).

[23] Ibid., p. 42.

apperceptive character of thinking enables us to reflect on our capacity for discursive cognition, to make explicit for ourselves that which is ordinarily only an implicit component of thought.[24]

Crucially, the critique of theoretical reason proceeds only against the background of reflection on our cognitive capacities, but reflection does not by itself amount to critique: in Henrich's words, the self-cognition achieved through reflection "is a source, not an achievement, of philosophical insight."[25] One of the aims of "philosophical investigation"—by which Henrich chiefly has in mind *critique*—is to discover and articulate connections and priority relations among the various facts and principles that govern and constitute our cognitive faculties, principles of which we are, in pure apperception, ordinarily only implicitly aware. For example, we might be implicitly aware that we can perform a certain cognitive act only on the condition of our performing a deeper or more basic one. It would be one of the tasks of a philosophical investigation to articulate and explain the structure of that relationship. But the pathway from implicit awareness to systematic philosophical inquiry will be mediated by reflection. Consider, for example, Kant's argument from the B-edition Transcendental Deduction that analysis "presupposes" synthesis, in that the analysis of a unified representation would not be possible unless the subject constituted the unity of that representation through an act of synthesis. Philosophical investigation enables us to articulate the fundamentality of synthesis with respect to analysis. However, this argument itself relies on an item of self-cognition disclosed through

[24] While Henrich stops short of *identifying* reflection with a kind of apperception, the connection to apperception is not lost on him: "the awareness 'I think' is precisely the self-consciousness that can be attached to natural and spontaneous reflection. And it is, in addition, the self-consciousness that can accompany every kind of reflection, regardless of the field of its employment" ("Kant's Notion of a Deduction and the Methodological Background of the First *Critique*," p. 45). What I believe Henrich means to say is that reflection on our cognitive capacities is itself a mode of pure apperception. Evidence for my reading of Henrich is found in the fact that Henrich proceeds to characterize the various respects in which reflection and pure apperception are formally alike, and supports this with a passage from the *Anthropology* where Kant refers to pure apperception as 'the reflected I' (p. 45).

[25] "Kant's Notion of a Deduction and the Methodological Background of the First *Critique*," p. 42.

reflection, namely, that our representing something as unified involves our "holding together" a multiplicity of representations actively in thought.[26]

In a more recent article, Houston Smit expands upon the basic ideas presented by Henrich, rendering the idea of critique's reliance on reflection more precise, as well as providing what is to my mind a much more compelling textual case that this was in fact Kant's view. Like Henrich, Smit holds that reflection is grounded in the pure apperception of the mind's activities.[27] And like Henrich, Smit holds that the critique of theoretical reason is an articulation and systematization of activities and principles revealed through reflection. Thus, a shared feature of their accounts is that reflection gives us a foothold into the operations of our cognitive faculties without which critique would be impossible. However, Smit expands on Henrich's account by highlighting a mode of reflection—*transcendental* reflection—that specifically discloses the conditions of our cognition of *objects*.[28] For Smit, this foothold enables us to isolate and articulate the structure of the formal conditions of objective cognition. Moreover, Smit is more explicit than Henrich about reflection's basis in pure apperception: for Smit, as for myself, it is in being conscious of "the identity of one's act of thinking and of the functions that one employs in thinking" that one is conscious of one's subjectivity.[29] In particular, Smit claims that pure apperception is a constitutive feature of cognition, and that in the pure

[26] Cf. Henrich ("Kant's Notion of a Deduction and the Methodological Background of the First *Critique*," pp. 43-44). For Kant's argument concerning the priority of synthesis, see KrV B130. We should expect that the boundary between reflection and investigation will be somewhat vague: as soon as we start reflecting on our cognitive capacities for the purposes of critique, our investigation has in some sense begun. The fact that Henrich writes as if there is a sharp boundary between reflection and investigation is, I believe, a consequence of the fact that he does not distinguish, as I do, ordinary pure apperception from the intentional activity I have been calling "reflection." (Of course, the boundary between ordinary pure apperception and reflection might itself be somewhat vague.)

[27] "The Role of Reflection in Kant's *Critique of Pure Reason*," p. 210.

[28] Ibid., p. 216.

[29] Ibid., p. 208.

apperception of such cognition we are conscious of the categories as specifying the "form of our cognition of objects."[30]

Smit shores up the textual case by pointing to the connection Kant makes between the task of a critique of pure theoretical reason and what Kant calls "transcendental philosophy." Kant writes,

> I call all cognition transcendental that is occupied not so much with objects but rather with our mode of cognition of objects insofar as this is to be possible *a priori*. A **system** of such concepts would be called **transcendental philosophy**. (KrV A11-12/B25)[31]

By such a "system" Kant has in mind the totality of *a priori* theoretical knowledge that is possible for us. The task of a critique of pure theoretical reason is not to develop such a system but instead to prepare for its possibility by determining the scope of reason's *a priori* cognitive powers. Thus, while critique is not, on this definition, transcendental philosophy, it is a form of transcendental cognition, since its concern is the possibility of *a priori* cognition.

This conception of critique presumes that we can isolate the role of our cognitive capacities in a determination of the *a priori* conditions of theoretical cognition. Smit relates this presumption to the distinctive capacity Kant calls "transcendental" (as opposed to "logical") reflection. The distinction between these two modes of reflection tracks the distinction Kant makes between transcendental and general logic. We saw that pure general logic is the study of the principles constitutive of thought, without consideration of how thought relates to an object. *Transcendental* logic likewise concerns the principles of thought, but differs from general logic in being specifically concerned with the *a priori* conditions of the relationship of thoughts to their objects (KrV A56/B80-81). Thus, the critique of theoretical reason is, in part, an exercise in transcendental logic.

Smit marshals persuasive textual evidence that Kant held that a science of general logic is made possible through our reflecting on the general logical form of our thought. (Following Smit, I will call

---

[30] Ibid., p. 208.

[31] Cf. KrV A56/B80-81.

this procedure 'logical reflection'; it is a species of the more general activity of reflection described above.) In the *Jäsche Logic,* Kant claims that we can achieve "insight" into the rules of general logic *a priori*, precisely because such rules "contain merely the conditions for the use of the understanding in general" (JL 9:12). Such rules are derived from "the necessary use of the understanding, which *one finds in oneself apart from all psychology*" (JL 9:14, my emphasis). The science of general logic therefore constitutes "a self-cognition [*Selbsterkenntnis*] of the understanding and of reason, not as to their faculties in regard to objects, however, but merely as to form" (JL 9:14). The notion that this science involves "find[ing] in oneself" the necessary forms of the understanding and reason might be taken to imply that it is a science of introspection, which for Kant would mean that general logic has its basis in empirical apperception. But such a reading is refuted by Kant's claim that the self-cognition of general logic does not belong to psychology. In a passage in the *Anthropology* in which Kant once again contrasts pure apperception ("consciousness of understanding") with empirical apperception ("consciousness of inner sense"), Kant explicitly identifies logic with the study of what is given in pure apperception and psychology with what is given in inner sense (An 7:134n).[32] Thus, when Kant claims in the *Jäsche Logic* that general logic constitutes a kind of self-cognition, *distinct from psychology*, in which one finds "in oneself" the necessary rules of the understanding, he has in mind specifically logical *reflection*, which is made possible by the pure apperception of the understanding in its logical use.[33]

The view that emerges is that logical reflection enables us to cognize those rules of thought that constitute the general logical form of our judgments.[34] As we saw in the previous chapter, the

[32] Strictly speaking, Kant asserts that logic is an investigation "according to what intellectual consciousness suggests" (An 7:134n). But given the context of this passage, it is obvious that 'intellectual consciousness' refers to pure apperception.

[33] Cf. Smit ("The Role of Reflection in Kant's *Critique of Pure Reason*," pp. 213-215).

[34] Smit draws a sharper distinction than I would like to between the "normatively necessary laws" of pure general logic and the "form" of our thinking, claiming that our pure apperception of logical form is an awareness not only of *how our thought is in fact formally logically structured* but moreover an awareness that our thinking is normatively governed by certain laws. Smit seems to hold that the former could not provide an awareness of normative constraint, but merely a description of our thinking (ibid., p. 214). On the constitutivist reading that I favor, the pure apperception of general logical form just

basic task of the understanding is to unite representations according to certain logical functions of judgment. To cite one of Kant's own examples, the proposition "If there is perfect justice, then obstinate evil will be punished" is a unity of distinct representations (the judgments "there is perfect justice" and "obstinate evil will be punished") according to the function particular to what Kant calls "hypothetical" judgment (the functional relation expressed by the "if…then..." formulation). The various functions define the general logical forms that judgments can exhibit. In general logic we isolate these forms, making explicit for ourselves the general rules of thinking. This requires logical reflection, through which we make explicit for ourselves the general logical form of our thought.

In the Amphiboly chapter, Kant specifies that there are certain pairs of concepts, which he deems "concepts of reflection," through which we represent the logical form of our thought: identity and difference, agreement and opposition, the inner and the outer, and matter and form (KrV A262/B318-A266/B322). For example, we deploy the concepts of identity and difference in our representation of what Kant calls the relation of "quantity" of a given judgment—our representation of whether it has the form of a *universal* judgment ("All Xs are Ys"), a *particular* judgment ("Some Xs are Ys"), or a *singular* judgment ("This X is Y") (A70/B95-A71/B96). Likewise, we deploy the concepts of agreement and opposition in our representation of the relation of "quality," whether *affirmative* ("Xs are Ys"), *negative* ("no Xs are Ys"), or *infinite* ("Xs are non-Ys") (ibid.).[35] We can therefore think of logical reflection as an activity guided by these "concepts of reflection," whereby the general logical form of thought becomes an explicit item of self-cognition. Moreover, logical inquiry would be impossible without this epistemic foothold into our own capacity for thought. Although Smit does not say as much, this suggests that the so-called "metaphysical" deduction of the categories, where

is consciousness of the general logical laws that we self-consciously deploy in thought and that constitute the logical form of our thinking. This does not entail that we can never err in our deployment of these rules.

[35] The notion of a "concept of reflection" adds a layer of description to the account provided in chapter one of the pure apperception of logical form. There I said that our consciousness of the "functions" of unity, most notably the twelve logical functions of judgment, is what makes possible an analysis of our judgments in terms of their logical form.

Kant first isolates the various logical forms and their associated categories in preparation for the "transcendental" deduction, begins as a general logical investigation undergirded by logical reflection. That section commences with Kant announcing that "we find" twelve distinct functions of judgment. We arrive at the table of judgments by "abstract[ing] from all content of a judgment in general, and attend[ing] only to the mere form of the understanding" (A70/B95). I want to suggest that this process of "abstracting" and "attending" involves deliberate logical reflection, and that Kant expects that the reader would likewise "find" within herself precisely these twelve functions of judgment, if she were to reflect on the general logical form of her thinking.[36]

Smit proceeds to claim that just as logical reflection is indispensable to general logic, so transcendental reflection is indispensable to transcendental logic and therefore the critique of pure theoretical reason. I said that the task of transcendental logic is to isolate the *a priori* contribution of our cognitive faculties to the possibility of our cognition of objects. Through mere logical reflection the thinker can determine that her judgment has, e.g., categorical form ("S is P"), but since general logic abstracts from all considerations of the content of cognition, this activity does not imply any cognizance on the part of the thinker that an *object* is represented by the subject ("S") of her judgment—that is, that some*thing* is being represented under the category of *substance*—or that the predicate of her judgment ("is P") represents one of the object's *accidents*, or properties. The self-consciousness implicit in taking our thought to be objective, to have content, therefore cannot be explained on the basis of our capacity for (mere) logical reflection. In brief, Kant claims in the Amphiboly chapter that we regard our thinking as having *content* by being conscious of it *as thinking* that relates *to sensibility* (A262/B318-A263/B319). This in turn requires cognizance on the part of the

[36] This should not be taken to imply that general logical inquiry can be undertaken independently of inquiry into the *a priori* conditions of the possibility of our cognition of objects—what Kant calls "transcendental" cognition. The characterization of general logic as freestanding with respect to transcendental logic has been disputed by Longuenesse (*Kant and the Capacity to Judge*, p. 76). I do not take up this question in this dissertation.

subject of the cognitive power from which a given representation originates—whether, for example, it originates in sensibility or understanding. Kant labels the activity through which we achieve this cognizance "transcendental" reflection (ibid.).

We saw that mere logical reflection consists in the subject's deliberately drawing into focus the general logical form of her thought, in abstraction from the conditions under which thought applies to an object. There is a parallel role for transcendental reflection. By reflecting on the transcendental conditions of cognition (as opposed to mere logical form), the subject makes her own capacity for objective cognition an explicit item of self-cognition. Such reflection is based in the pure apperception of the specifically *cognitive* activity of pure understanding—of the understanding in its subsumption of sensible intuitions under pure concepts (i.e., the categories). And while pure apperception is a constitutive feature of ordinary objective cognition, rational subjects can deliberately undertake to *reflect on* the transcendental conditions of cognition. By attending carefully to what is "given," so to speak, in the pure apperception of cognition, in particular to the various *a priori* functions of the understanding with respect to our cognition of objects, we can make explicit for ourselves certain formal conditions that any possible object of our cognition must meet. Ultimately, and after careful investigation, we can establish that such objects necessarily conform to the principles of pure understanding.[37]

This is, admittedly, a highly general account of the role of transcendental reflection with respect to Kant's program in the first *Critique*. I have mostly elided any distinction between the so-called "metaphysical" and transcendental deduction of the categories, as well as any distinction between these deductions and later chapters of the Transcendental Analytic devoted to the schematized versions of the categories and the principles of pure understanding. Moreover, a fuller treatment of how, on Kant's view, subjects come to regard their representations as having objective

[37] Cf. Smit ("The Role of Reflection in Kant's *Critique of Pure Reason*," p. 216).

significance lies beyond the scope of this dissertation.[38] For our purposes, the important takeaway is that transcendental reflection constitutes a source of self-cognition that the critique of pure theoretical reason draws upon in its delineation of the sources and limits of theoretical cognition. To forestall any misunderstanding, let me again state that this does not entail that the aims of critique can be achieved *merely* through reflection. Such a view overstates the exactness and systematicity of the self-cognition that reflection affords us, rendering the *Critique of Pure Reason* otiose. Nevertheless, the view does entail that reflection constitutes an indispensable starting point of the critical *method*. And if that is right, then at the heart of Kant's attempted critique of pure theoretical reason lies a *methodological assumption* that rational subjects are capable of transcendental reflection.

Support for this last claim is found in methodological remarks made early on in both editions of the first *Critique*. In the A-edition Preface, shortly after setting forth the idea of a critique of pure reason outlined in the previous section, Kant attempts to assuage concerns the reader might have about the possibility of such a critique. The tenor of this brief discussion suggests that the concerns he has in mind have a specifically epistemological cast, as if, according to the underlying worry, establishing the scope and limits of reason's principles lies beyond reason's ken. In response to this concern, he writes:

> I have to do merely with reason itself and its pure thinking; to gain exhaustive acquaintance with them I need not seek far beyond myself, because it is in myself that I encounter them… (Axiv, my emphasis)

Later, in a passage that appears in both the A- and B-edition Introduction, Kant returns to the topic of the nature and possibility of a critique of pure reason. He writes:

[38] Furthermore, I do not mean to suggest that subjects' capacity to distinguish the understanding and sensibility as heterogeneous sources of cognition suffices to explain how on Kant's view subjects come to regard their thinking as having objective content. Any remotely satisfactory account of this aspect of Kant's thought will have to address in detail the role of the various functions of judgment vis-à-vis the categories in constituting what Kant calls the "objective" unity of apperception (B141). Cf. Longuenesse (*Kant and the Capacity to Judge,* pp. 52-56, Part Two, and especially Part Three, where in each case, the relation between category and schema is taken into consideration to ground the objectivity of the relevant category) as well as Boyle ("Kant's Categories as Concepts of Reflection" (unpublished)).

> For that [a critique of pure reason leading to a complete system of pure reason] should be possible…can be assessed in advance from the fact that our object is not the nature of things, which is inexhaustible, but the understanding, which judges about the nature of things, and this in turn only in regard to its *a priori* cognition, the supply of which, since we do not need to search for it externally, cannot be hidden from us… (A12-13/B26, my emphasis)

Given the context of these remarks, it is clear that Kant believes that rational subjects have a capacity for self-cognition of the sort that enables critique. Both passages employ spatial metaphors, with the relevant sort of cognition being achieved by the subject looking "within" herself. The second passage appears to make the even stronger claim that, because the topic of investigation is the subject's own understanding and therefore not something one must "search for…externally," it is *necessarily* of the sort that the subject could achieve cognition of it sufficient for its critical investigation.

Given what has been said so far, it is natural to suppose that Kant is referring in these passages to our capacity for transcendental reflection and its essential role within a possible critique of pure theoretical reason. But the spatial (and, in the latter, visual) metaphors that Kant employs are liable to mislead the reader, for they could be taken to suggest that reflection involves a kind of introspection, and that central to the critical enterprise is the assumption that the mind is in all ways transparent to itself. I will argue that this is wrong on both counts.

The discussion thus far has supposed that reflection has its basis in pure apperception, not introspection. On this view, reflection is itself a form of pure apperception: it is the act of pure apperception whereby I attend to those features of my thought that I apperceive, such as its logical form. We have already seen some textual evidence for this claim, specifically Kant's assertion that general logic is the study of those rules of the understanding that "one finds in oneself apart from all psychology" (JL 9:14). I will now present additional textual support for this claim.

The most direct textual support comes from Kant's *Anthropology*, where Kant appears in several places simply to *identify* reflection with pure apperception. Here I must remind the reader of a terminological ambiguity. We saw that in the Amphiboly chapter of the first *Critique*, Kant

characterizes reflection as, among other things, "the state of mind in which we first prepare ourselves to find out the subjective conditions under which we can arrive at concepts" (A260/B316). On the basis of this quotation, and in order to regiment the notion of reflection, I stipulated that reflection is a voluntary undertaking for the purpose of achieving explicit self-cognition. But Kant does not carefully distinguish reflection *qua* voluntary activity from pure apperception more generally. Indeed, in both the *Anthropology* and the Amphiboly chapter he appears at times to use the term 'reflection' to refer to the pure apperception that is an involuntary, constitutive feature of all thinking. This suggests that Kant did not regard my distinction to be especially significant. This in turn strongly suggests that reflection *qua* voluntary activity is not a form of introspection but rather a form of pure apperception.

In a footnote in the *Anthropology* that we have already examined,[39] Kant refers to the self-consciousness "of reflection," the consciousness of our "inner activity" or "spontaneity" (An 7:134n). He then characterizes this consciousness as "consciousness of understanding," which he alternately labels "pure apperception" (ibid.). A few pages later, in a section titled "On sensibility in contrast to understanding," he remarks,

> [C]ognition (since it rests on judgments) requires reflection (*reflexio*), and consequently consciousness of activity in combining the manifold of ideas according to a rule of the unity of the manifold… Discursive consciousness (pure apperception of one's mental activity) is simple. The "I" of reflection contains no manifold in itself and is always one and the same in every judgment, because it is merely the formal element of consciousness. (An 7:141)

As I understand this passage, Kant is claiming that judgment is necessarily accompanied by consciousness of the activity of judging, which he deems "reflection." But the generic form of this awareness is articulated in terms of the idea of "consciousness of activity in combining [a] manifold…according to a rule of the unity of the manifold" (ibid.). In other words, reflection constitutes an awareness of our combining representations according to a rule specifying the formal

[39] See footnote 32.

relation in which those representations are made to stand (e.g., categorical or hypothetical form, etc.). But this awareness is precisely what Kant refers to here and elsewhere as "pure apperception," described in the passage above as a universal "formal element" of discursive consciousness. This implies that when Kant uses the term 'reflection' to refer to the self-consciousness of a rational activity, he is either using this term as a synonym for pure apperception (as in the expression "'I' of reflection") or pointing to a distinction *within* pure apperception (as when he distinguishes logical and transcendental reflection).[40]

We might nevertheless worry that when Kant refers, in the Amphiboly chapter of the first *Critique*, to a power of reflection through which we can identify the "subjective conditions" of our capacity for objective cognition (KrV A260/B316ff.), he has in mind a capacity that is not grounded in pure apperception. But this concern is mitigated by Kant's assertion that transcendental reflection is a basic requirement of all judgment (A261/B317). Since Kant lists pure and empirical apperception as the two most basic forms of self-consciousness, we can assume at least that transcendental reflection is equivalent to, or has its basis in, some type of either pure or empirical apperception. Of those options, only pure apperception is regarded by Kant as a necessary and constitutive feature of all judgment. This suggests that the very same formal structure of self-consciousness expressed by "I think" (i.e., pure apperception) is exhibited in transcendental reflection. To suppose otherwise is to suppose that in transcendental reflection we make our various cognitive powers, as well as their co-operation in empirical cognition, items of *empirical* as opposed to pure apperception—or, putting the point more generally, that the self-awareness of reflection has the same form as our cognition of objects. Given the indispensability of transcendental reflection to the critique of pure reason, this in

[40] Another possibility suggests itself: that at various places Kant is implying a connection between reflection as pure apperception and reflection as concept formation (universalization) from given representations. Such a connection is suggested, I believe, by Longuenesse (*Kant and the Capacity to Judge*) and positively defended by Boyle ("Kant's Categories as Concepts of Reflection" (unpublished)).

turn implies that the project of critique is essentially a form of empirical cognition. But as we saw above, Kant understands critique to be a mode of the self-cognition of reason that is neither a kind of empirical cognition nor a kind of conceptual analysis. Together with Kant's insistence that a capacity for self-cognition grounds the possibility of critique, this implies that we achieve this self-cognition through pure apperception.

By making explicit that reflection is fundamentally the *pure apperception* of a particular rational capacity (undertaken deliberately, according to my stipulation), we are able to respond to the worry that the project of critique rests on an assumption of mental transparency. In the methodological comments quoted above, Kant is referring primarily to our pure apperceptive awareness of the *a priori* functions of the understanding in the cognition of objects. Kant believes we have this awareness because he holds that the understanding is among our essentially self-conscious, rational capacities.[41] But from the fact that our *rational* capacities are essentially self-conscious, it does not follow that the mind is in *every respect* transparent to itself. While some commentators have interpreted the doctrine of the analytical unity of apperception as a transparency thesis,[42] Kant certainly allows for the possibility of representations to which the representation "I think" could not attach; such representations, we saw, "would be nothing for me" (KrV B132). Moreover, Kant claims that representation "with consciousness" is a subspecies of representation more generally (A320/B376), and his *Anthropology* contains an extended discussion of "representations that we have without being conscious of them" (An 7:135-137).

I will close this section by remarking upon the relationship of the doctrine of pure apperception to critique's reliance on reflection. Drawing on the work of Henrich and Smit, I have

[41] In other words, Kant subscribes to the following restricted transparency thesis: a rational subject is aware, through pure apperception, of the function of the *internal principle* of the rational activity in which she is engaged. This will become relevant in chapter four.

[42] Cf. Carruthers (*The Opacity of Mind: An Integrative Theory of Self-Knowledge*, p. 27).

argued that critique proceeds from claims about the nature of our rational capacities, including our capacity for discursive cognition, that are items of reflective self-cognition. I have also argued that the activity of reflection is a mode of pure apperception. On my account, pure apperception consists in our consciousness of ourselves as agents of combination according to certain functions of unity among our representations. These "functions" include the pure cognitive functions of the understanding, that is, the subsumption of sensible intuitions under categorial concepts. Thus, if pure apperception is a constitutive feature of our various discursive capacities, including our transcendental capacities, then these very capacities are available for reflective articulation in a critique of their origin and scope. In this way, the *fact* (assuming it is a fact) of pure apperception enables critique's methodological reliance on reflection.

Nevertheless, I believe that the *doctrine* of pure apperception—that is, the theory comprising Kant's official claims about the nature of pure self-consciousness and its role in our representational economy—is by Kant's own lights ultimately justified through reflection on our own case. Kant invokes the doctrine of pure apperception in order to account for our capacity to represent an object's unity, a capacity that in the previous chapter I argued cannot be explained on the basis of a higher-order representation. That capacity was explained in connection with the principle of the analytical unity of apperception, which specified a condition that all representations must meet in order to be something "for me," their subject: all representations must be such that they could figure as the content of a possible judgment of the form "I think x," that is, one in which I represent myself as the self-identical thinker of that content. In the Transcendental Deduction of the Categories, Kant claims that we can represent the unity of an object through our awareness of how the particular constituent representations that jointly constitute our representation of that object are related. But this amounts to an awareness of the *function* according to which we actively relate together such representations in thought, and thereby of ourselves as the "agents" that bring about the unity associated with that

function. This consciousness, which Kant labels "pure apperception," is precisely what the generic representation "I think" expresses. The Transcendental Deduction of the Categories therefore prescinds from our capacity to represent objective unity to a capacity for pure apperception. But as I stated at the outset of this chapter, the claim that we possess a capacity for non-introspective (i.e., non-empirical) self-consciousness, and that this self-consciousness consists in the consciousness of ourselves as agents of combination, can itself be questioned. Moreover, Kant does not provide an argument for the reality of pure apperception apart from his invoking it in the above accounts. The best explanation for this, I believe, is that Kant expects the pure apperceptive character of thought to be available to the reader as an item of reflective self-cognition. In other words, in invoking the pure apperceptive character of thought in an explanation of our capacity for discursive cognition, Kant is implicitly appealing to the reader's own capacity for reflective self-cognition, and he expects the reader to assent to the explanations he provides because he takes himself to be citing features of cognition that are available for reflective articulation. This is precisely what we should expect if, as I have been arguing, critique relies upon reflection as a source of self-cognition. For the most fundamental item of self-cognition afforded to rational subjects by reflection, regardless of the particular discursive activity on which we are reflecting, is our status as *active* subjects who deploy principles in thought.

§2.3 Allow me to recap the main claims of this chapter so far. In §2.1, I presented an outline of how Kant conceives of the project of critique. There I said that because critique issues in synthetic *a priori* claims, its task necessitates a source of substantive, non-empirical self-cognition. In §2.2, I argued that reflection on our discursive capacities supplies the requisite cognition, and that such reflection is a mode of pure apperception. But for all that has been said so far, reflection could be merely an epistemic enabling condition for critique's operation; perhaps under different epistemic constraints, critique could be fulfilled without the aid of reflection. If that were the case, then reflection would be

materially indispensable to critique but nonetheless *"conceptually" incidental*. In this section I will argue that critique is a reflective/apperceptive procedure *in conception*. In particular, I will argue that critique's reliance on reflection is an expression of the fact that Kant conceives of critique as exhibiting the autonomy of human reason. What follows is an addendum to, not a criticism of, the account of reflection offered in the previous section. The material of this section will provide necessary background to chapter four, where I will argue that the critique of practical reason likewise presupposes a capacity for reflection grounded in a commitment to the autonomy of reason.

By making explicit that Kant conceives of critique as an expression of our rational autonomy, we can also form the beginning of a response to—and a reframing of—epistemological and methodological misgivings about critique's reliance on reflection. The *Critique of Pure Reason* makes a series of claims about the psychology of any finite rational subject. Modern readers are liable to worry that such claims are at best just articulations of "how things seem to us" and thus involve a scientifically disreputable introspectionism. Sure, these critics might admit, by the theory's own lights its method isn't introspectionist in the sense of resting on the deliverances of specifically *empirical* apperception; nonetheless, its confident assertion of psychological capacities in place of any attempt to "operationalize" them for empirical verification has all the trappings of a discredited introspectionism and thus dooms it from the start. Against this, we can say that critique's reliance on reflection is grounded in reason's presumption of its own normative autonomy. Thus, its claims constitute a form of explanation oriented around a particular normative demand. Moreover, it does not oppose empirically-tested psychology so much as situate this domain of inquiry within a larger critical framework. While such a response hardly settles the debate about method, it at least shifts the terms, for the relevant question now becomes whether the assumption of rational autonomy is proper to, and perhaps a necessary feature of, philosophical theorizing.

Recall that Kant conceives of the critique of pure reason as a "court of justice" instituted by reason itself, according to "[reason's] own eternal and unchangeable laws" (Axi-xii). In §2.1, I cited this as evidence that Kant conceives of critique on the model of the notion of rational autonomy that he articulates in his moral philosophy. However, it must be acknowledged that Kant never uses the term 'autonomy' or its cognates to describe the critical method. Indeed, 'autonomy' is deployed by Kant principally as a technical term that denotes a feature of the *wills* of moral agents. Moreover, in the context of such usage, autonomy implies "transcendental" freedom of the will, negatively defined as freedom from determination by natural causes (KpV 5:97).[43] Kant identifies the will with practical reason (G 4:412) and holds that this capacity is associated with a power of *choice* (*Willkür*) (Rel 6:25; MS 6:213). On this construal, autonomy is regarded as the distinctive capacity to act on the basis of principles of *pure* practical reason, and implies a power to choose to act on such principles independently of natural causal determination.

To be sure, I am not asserting that Kant conceives of critique as an exercise of practical reason or associates it with the power of transcendentally free *Willkür*. Thus, I do not hold that Kant conceives of critique as "autonomous" in exactly the above sense. However, I will argue that Kant does regard critique as sharing with autonomous agency at least these two interrelated features: the exhibition of a capacity for *rational self-legislation* and, essential to that capacity, a particular *normative status*.

Again, Kant undertakes in the first *Critique* to subject reason's claim to synthetic, *a priori* knowledge to "[reason's] own eternal and unchangeable laws" (Axi-xii). This *self-governing* or *self-legislative* feature of critique—the idea of critique as operating from *reason's own principles*—quite clearly seems to share something in common with Kant's conception of autonomous agency, which he routinely describes as proceeding from practical reason's own pure principles. In the *Groundwork*, this

[43] Cf. KrV A533/B561.

idea is reflected in the notion of the moral agent as a "lawgiver" who stands under laws "given by himself" (G 4:432), a point that is repeated in the second *Critique*'s discussion of autonomy (KpV 5:33). In these works, Kant contrasts autonomy with heteronomy, and the conception of practical reason as lawgiving is usually contrasted with the idea of a will whose ends are determined by some ultimate material aim—the agent's own happiness, for example, or even the (average or total) collective happiness of all rational beings (G 4:442ff.). But Kant explicitly includes rationalist perfectionism and divine command theory among the moral theories he regards as heteronomous (ibid.). Kant is clear, in other words, that the ultimate ground of moral action must be the agent's own pure practical reason—not some universal desire for happiness, not some external moral order to which we must submit, and not even the will of God.

But if reason is the ultimate source of the principles by which it regulates itself, then reason is itself sovereign with respect to any putative "external" normative standard. To say this is to make salient something that is contained in the idea of reason as a rational self-legislator. To be the source of the principles to which one must conform is to have a particular *normative status*. It is to be free from external normative constraint. In the *Groundwork*, Kant makes this status explicit in his discussion of the moral agent as giving law through his or her own pure practical reason:

> [T]he will is not merely subject to the law but subject to it in such a way that it must be viewed as also giving the law to itself and just because of this as first subject to the law (of which it can regard itself as author)…
>
> If we look back upon all previous efforts that have ever been made to discover the principle of morality, we need not wonder why all of them had to fail. It was seen that the human being is bound to laws by his duty, but it never occurred to them that he is subject *only to laws given by himself but still universal* and that he is bound only to act in conformity with his own will… (G 4:431-2)

Kant proceeds to call the idea that the will is obligated only by its own laws "the principle of the **autonomy** of the will" (G 4:432). Thus, central to Kant's notion of (practical) autonomy is the conception of pure (practical) reason as unconstrained by external normative standards in its capacity as the ultimate source or "legislator" of its own normative principles.

But the basic notion of a mode of rational self-legislation that is free from external normative constraint is not restricted to Kant's discussion of pure practical reason; indeed, as we have seen, it is already present in his discussion of the concept of a critique of pure reason. In Kant's metaphor, reason in its self-critique is not just operating according to its own laws and principles, but explicitly presiding over a "court of justice" to which reason submits its claims. According to Kant's metaphor, pure reason is the ultimate "authority," so to speak, with respect to the scope and legitimacy of its own principles; and this implies that pure reason in its self-critical undertaking is likewise normatively unconstrained by external standards. This claim is underscored within Kant's discussion by his several allusions to the sovereignty of reason being a chief principle of the Enlightenment. The putative "external" standard with which he contrasts the idea of critique appears to include any sort of "received"—and therefore uncontested and uncritically accepted—claims to knowledge. He asserts that the "ripened power of judgment" of his age, "which will no longer be put off with illusory knowledge," positively *demands* that reason should undertake a critique of itself. And in a footnote to this discussion, Kant explicitly lists religion and royal decree as among the former authorities that are now subject to (and thus should no longer be regarded as supplying the standards for) rational scrutiny:

> Our age is the genuine age of criticism, to which everything must submit. **Religion** through its **holiness** and **legislation** through its **majesty** commonly seek to exempt themselves from it. But in this way they excite a just suspicion against themselves, and cannot lay claim to that unfeigned respect that reason grants only to that which has been able to withstand its free and public examination. (KrV Axin)

At the risk of belaboring the point, I submit that what Kant refers to in the above passage as 'religion' and 'legislation' occupy the same functional role as did the various possible foundations for heteronomous moral theorizing that Kant cited in the *Groundwork*. In both cases, the authority of some putative source of a rational requirement is contested on the grounds that reason is the source of its own principles and thus "free" with respect to the authority in question.

In the *Groundwork*, Kant claims that the presupposition of rational autonomy (broadly construed) is an essential feature of reason *as such*.

> [O]ne cannot possibly think of a reason that would consciously receive direction from any other quarter with respect to its judgments, since the subject would then attribute the determination of his judgment not to his reason but to an impulse. Reason must regard itself as the author of its principles, independently of alien influence… (G 4:448)

In the context of this passage, the claim that *reason* must regard itself as the author of its principles functions as the major premise of a syllogism whose conclusion is that, necessarily, every rational agent must regard his or her practical reason as autonomous.[44] (The unstated minor premise is that practical reason is an employment of reason.) Kant is therefore claiming that reason *in general*—not *practical* reason in particular—presupposes its own status as a rational self-legislator. And while Kant does not deny that our judgment sometimes succumbs to non-rational interference, he nevertheless denies that reason can ever self-consciously make a determination on grounds other than its own principles.

Kant's reference to an "impulse" that determines judgment suggests that he has in mind principally the non-rational influence of sensible desire on practical deliberation—a topic we will address in later chapters. But Kant is also implying, for example, that a rational subject engaged in theoretical deliberation cannot self-consciously abrogate her status as the epistemic agent ultimately responsible for deciding what to believe, and therefore someone who must judge according to her own theoretical standards. This does not mean that she cannot, for example, lend weight to expert testimony, and even (in the colloquial sense) "completely defer" to others. Nor does it mean that she is immune to bias, irrational prejudice, defective reasoning, and so forth. However, it does entail that she cannot *consciously* form theoretical judgments on the basis of considerations that are arbitrary according to the standards of her own theoretical reason. In such a case, she could not regard her judgment as supported by considerations of the sort she could take up in an act of conscious

[44] I will address this claim in chapter four.

deliberation. But that is just to say that she could not regard her judgment as meeting the necessary conditions of its being her judgment, and so could not regard it as her judgment at all.

The foregoing account of the nature of reason sheds considerable light on the first *Critique*'s assertion of a "demand" for a critique of pure reason, which Kant credits to the "ripened power of judgment" of his age (KrV Axi). First, it traces this demand to a constitutive feature of reason: reason's necessary presumption of its own autonomy. Moreover, it suggests that reason's demand for self-critique is concomitant with reason's becoming conscious of—and consequently rejecting—its previous reliance on standards for judgment that are arbitrary from the standpoint of reason itself. This is because, if reason cannot *consciously* make determinations on the basis of an uncritically accepted "authority," so reason cannot sustain any such acceptance of which it *becomes conscious*. Thus, when Kant says that his age "will no longer be put off with illusory knowledge," he is describing what he regards as a necessary consequence of rational subjects achieving awareness of their dogmatic reliance on religion, social convention, and other external normative standards in the determination of their beliefs (KrV Axi, my emphasis).

Within the development of this "age of criticism," the critique of pure reason occupies a special role. Regardless of its object, the general aim of criticism in the broad sense is to determine, according to rational principles, the legitimacy of some normative standard—for example, the legitimacy of an evidential standard in a court of law, or the legitimacy of some legislative body. But any such determination will of course presume the legitimacy of the principles it applies. As self-conscious beings, subjects engaged in criticism in this broad sense can achieve explicit awareness of their application of principles and subject certain of those principles to criticism. Such is the task of the critique of pure reason, which is reason's *self*-critique. Inasmuch as criticism in general seeks to

determine the legitimacy of some standard, the fulfillment of reason's self-critique necessitates a determination of the legitimacy of that subset of our rational principles whose legitimacy is in dispute.[45]

It is important to keep in mind, however, that this "dispute" occurs within reason itself, and therefore *not* on the basis of some external-to-reason "check" on the rational authority of reason's own principles. Consider, by way of illustration, the second analogy of experience, the "principle of temporal sequence according to the law of causality" (KrV B232). In the B edition of the first *Critique*, this principle is stated as follows: "All alterations occur in accordance with the law of cause and effect" (ibid.).[46] This is a principle of pure understanding and therefore, according to Kant, a principle whose application is necessary to achieve objective cognition (A159/B198). A major task of the first *Critique* is to validate this principle as one that in fact applies to empirical objects. However, reason's capacity for critical self-scrutiny is a source of doubt about the legitimacy of this principle, in two ways. First, since the principle of cause and effect is a synthetic, *a priori* principle, a suspicion arises as to what (if anything) licenses its application to empirically-given objects. (This is just an instance of the more general problem of synthetic, *a priori* judgment, which I won't rehearse here.) Second, in the exposition of the third antinomy Kant argues that the unrestricted application of this principle gives rise to an apparent opposition *within* reason, because seemingly equally-valid arguments yield contradictory conclusions (A444/B472ff.). This is one among a set of "antinomies" within reason itself. Kant makes an oblique reference to the antinomies in his A-edition remarks on the idea of a critique of pure reason.

---

[45] Given that critique involves reason's determination of the legitimacy of its own principles, two questions arise: (1) What does it mean for the legitimacy of some such principle to be "in dispute"? (2) Does reason presume the legitimacy of the very principles whose legitimacy it means to establish (and is this not objectionably circular)? With respect to (1), we shall see that reason is itself the source of doubts about the legitimacy of its principles. Thus, the relevant "dispute" is located within reason itself. With respect to (2), we shall see that reason need *not* presume the legitimacy of the very principles it subjects to critique. Rather, on a plausible but charitable interpretation of Kant's attempt at a critique of pure reason, critique employs certain principles in a determination of the legitimacy of other principles. If that is correct, the principles that reason deploys in critique and the principles that reason submits to critique form disjoint sets. See footnote 47.

[46] In the A edition, the principle is called the "Principle of Generation" and is stated as follows: "Everything that happens (begins to be) presupposes something which it follows in accordance with a rule" (A 189).

There he claims that critique is necessary to "remov[e] all those errors that have so far put reason into dissension with itself in its nonexperiential use" (Axii). The point I wish to emphasize is that reason, even in self-critique, never abandons the presumption of its normative autonomy. The determination of the scope and legitimacy of its principles, as well as the resolution of its internal conflicts, are to be done "to reason's full satisfaction" (ibid., my emphasis).

The idea of a critique of pure reason, at least as Kant presents this notion in the A-edition Preface, gives rise to the worry that its method is circular. For how can reason critique its own principles except on the basis of its own principles? Against the worry that critique engages in circular reasoning, it is important to keep in mind that Kant is not critiquing reason in every aspect, but chiefly those principles of reason (including the understanding) that purport to yield synthetic, *a priori* cognition. For example, he makes no attempt to justify—i.e., validate in a self-critique—the principle of non-contradiction. Rather, he is attempting to demonstrate the possibility, but also the limits, of synthetic, *a priori* cognition. Thus, for example, he endeavors to critique the legitimate application of the concepts of cause and effect, in particular the principle of cause and effect cited above. We can therefore make a distinction between the principles that Kant *deploys in critique* and those that he *submits to critique*. We need not suppose that these principles are the same, or that Kant ever applies *in critique* those principles whose legitimacy is a topic of critical investigation.[47, 48]

---

[47] This is not to claim that Kant never does this. My point is only that the idea of critique, as presented, is a coherent one. Indeed, some commentators have worried that Kant's notion of "affection" by things-in-themselves illicitly smuggles in a notion of causation to which he is not entitled and which conflicts with his own views about the limitations of our knowledge. For a helpful discussion and defense of Kant on this point, see Longuenesse (*Kant and the Capacity to Judge*, p. 22 fn. 11). See also footnote 45.

[48] This raises the question of precisely *what* principles, in the broad sense of that term, are deployed in reason's self-critique. That is, what are the normative standards according to which Kant believes reason can determine the legitimacy of those principles it submits to critique? I claimed in §2.1 that among the principles Kant deploys are formal logical principles. We can think of the present section as arguing, in effect, that there is at least one additional "principle" of critique, namely, a commitment to critique as exhibiting a form of rational autonomy, normatively answerable only to pure reason, with an attendant presumption of a capacity for reflective self-cognition. But does Kant take these principles to be sufficient to the task of critique, or are there others? Unfortunately, Kant never precisely identifies the specific normative standards that regulate and constrain his project. Thus, in answering this question the best we can do is reconstruct the various arguments that constitute his attempt at a critique of pure reason, trying to identify along the way the normative standards

In light of this distinction, we can ask: What is the relationship of the subject to those principles she deploys *in critique*? What is the relationship of the subject to those principles she submits *to critique*? Drawing on my earlier conclusion that Kant conceives of critique as exhibiting the autonomy of reason, I will argue that in both cases the relationship is given by the form of pure apperception articulated in chapter one.

Consider first the principles that the subject deploys in critical investigation. As I noted in §2.1, such principles include the pure formal principles of general logic, including the law of non-contradiction.[49] This "law" will serve to illustrate the formal relationship the subject bears to the principles she deploys in critique. The first thing to note is that critique is a self-conscious procedure. For critique to exhibit the autonomy of reason in the sense adduced above, the subject must be able to apply the principle of non-contradiction compatibly with the self-conscious representation of her own reason *as autonomous*. That is, she must be able to represent her own power of reason as the source of the rules according to which she determines her judgment, and therefore as normatively unconstrained by putative "external" requirements. If critique weren't a self-conscious procedure, then its operation could not be seen as *meeting* the general critical demand to determine the rational legitimacy of normative standards according to reason's own principles. This is because the fulfillment of the self-critical task consists in the subject achieving explicit self-cognition, in particular, cognition of the limitations and legitimacy of various of her rational principles, *along with* recognition that this

that are operative in such arguments. As I noted in the previous footnote, some commentators have worried that Kant illicitly assumes a causal principle, but I have suggested that this need not be the case. I do not in this chapter try to provide an exhaustive list of the normative standards operative in critique.

[49] The law of non-contradiction originates in the formal logical reflection on what Kant calls the "quality" of our judgment according to the concepts of *agreement* and *opposition*. For example, an *affirmative* judgment contains a reflective (i.e., pure apperceptive) component according to which the subject and predicate are represented as in formal logical *agreement* (logical coherence) with each other. In the Amphiboly chapter, Kant argues that if the reflected representation of agreement and opposition consisted only in the representation of formal logical coherence or conflict, no representation of "opposition between realities," such as between equal and opposite vector forces, would be possible (KrV A264-5/B320-321). Cf. Longuenesse (*Kant and the Capacity to Judge*, Chapter 6, esp. pp. 136-139).

cognition is established not according to external standards, but by reason's own standards of critical self-appraisal.

Moreover, the requisite self-consciousness must have the form of *pure apperception*. This is because the subject must be able to represent herself as, in a certain sense, the *agent* of criticism. To see why, consider the inadequacy of the model of second-order representation in an account of the relevant form of self-consciousness. If the application of the principle of non-contradiction were represented by the subject in the form of an object of representation, i.e., as something distinguished in representation from her own subjectivity, then she would be representing the application of this principle in a way that alienates it from her own subjective, rational powers. In doing so, she could perhaps represent different sets of propositions as violating or conforming to this principle, but *as represented* the principle would have the form of an "external" normative demand, not one that issues from the subject's own reason.[50] To represent this principle as a requirement of her own reason, she must represent its application in a way that identifies her as the agent of its application. Put otherwise, she must, in self-consciously deploying the principle, represent its deployment as *something she does*. Furthermore, the representation of herself as *doing*, as the locus of agency, is not something over and above her representation of the application of the principle, but a constitutive feature of that very representation. But that is just to say that her self-consciousness has the form of pure apperception.

Consider next the principles that make up the topic of critical investigation—for example, the principle of cause and effect discussed above. To be sure, the subject does, in self-critique, consider this principle in the form of an object of representation. Indeed, subjecting a principle to critique involves taking up a perspective from which the rational legitimacy of the principle is not presumed,

[50] The qualifying term 'as represented' is necessary to the illustration of my point. Since the principle of non-contradiction is, according to Kant, a constitutive principle of discursive thought (see prior footnote), a subject could not discursively represent the application of this principle in the form of a second-order representation without at the same time at least implicitly applying that principle in such a way that constitutively identifies her as the agent of its application. Nevertheless, the principle *as represented* in my example cannot be regarded by her as issuing from her rational powers in the requisite sense.

even implicitly. This in turn involves representing this principle as alienated from the subjective, rational powers whose legitimacy is presumed *in the act of criticism*, that is, in the pure apperceptive application of principles such as the law of non-contradiction. Thus, for the purposes of self-critique, not every representation of the principle of cause and effect will involve the pure apperception of applying this principle in thought. In other words, not every representation of this principle will be such that it constitutively identifies the subject as the agent of thought whose act of thinking is the application of this principle. However, a critique of pure reason is a reflexive procedure that self-consciously identifies the principles it subjects to criticism as principles belonging to the subject's own pure powers of reason (including the understanding). For critique to produce in the subject the self-cognition expressed in the first person as "*My* pure principles have such-and-such scope of legitimate application"—for the subject to represent such principles as an expression of her rational autonomy—it must be possible for the subject, in critique, to identify the principles as hers in the requisite sense. That is, while she must be able to represent certain principles as alienated from her subjective, rational powers, she must also be able to represent them as an expression of her subjective, rational powers. And this requires her to be able to occupy a perspective in which she self-consciously applies, e.g., the principle of cause and effect in such a way that constitutively identifies her as the agent (thinker) actively applying this principle in thought—for example, in theoretical judgments concerning empirical objects. But that is just to say that she must be capable of self-conscious representations of this principle whose form is the pure apperception of thinking according to this principle.

The previous three paragraphs jointly entail that critique necessarily relies on reflection as a source of self-cognition. Critique is a self-conscious procedure in which the subject identifies certain principles as belonging to her own reason, in an attempt to determine their legitimacy according to reason's own standards. The fulfillment of the critical task therefore requires that the subject explicitly cognize both the identity of certain of her principles along with the fact that they *are in fact* principles

of her own reason. This is particularly clear in the case of those principles that she *submits* to critique. For example, in submitting to critique the principle of cause and effect, the subject must cognize the nature of that principle (i.e., what it asserts of the objects that fall within its scope), as well as the fact that the principle is one that she herself applies in thought. But according to the above argument, the sense in which the subject regards that principle as her own is explained in terms of her pure apperception of its deployment in ordinary thought. Thus, in order to submit this principle to a critique, the subject must identify a feature of her thought that is part of her pure apperceptive consciousness so that it can become an *explicit* item of self-cognition. But that is just to say that in critique she relies on reflection on her own discursive activity as a source of self-cognition. Somewhat less obviously, the critique of pure reason relies on reflection not just with respect to the identification of the principles that it *submits* to critique, but moreover with respect to its own activity. I argued that because critique is an autonomous procedure, the subject must regard herself as the "agent" of those principles she deploys in critique, and therefore that critique must have pure apperceptive form. But I also claimed that the fulfillment of critique includes the subject's recognition that the scope and legitimacy of certain of her principles have been determined *by reason's own standards*. Critique therefore involves an identification of its own standards *as standards that belong to the subject's own reason*, and since critique is a pure apperceptive procedure, this is again a form of reflective self-cognition.

Importantly, the foregoing conclusions are not just a reformulation of the earlier claim that critique necessarily relies on reflection as a source of self-cognition. As we saw, that assertion was compatible with reflection on our rational capacities being merely an epistemic enabling condition of critique's operation. Here we see that critique's reliance on reflection follows from Kant's *very conception of critique as an autonomous, reflexive procedure*.

Moreover, since Kant regards the demand for critique as issuing from reason's necessary presumption of its own normative authority, critique is a philosophical method whose purpose is to

fulfill a particular normative demand. I mentioned at the beginning of this section that modern readers are liable to worry that Kant's reliance on reflection amounts to a scientifically disreputable form of introspectionism. While it is not my aim to adjudicate this dispute, it bears mentioning that critique's reliance on reflection is not—or at least is not *only*—a consequence of an unduly permissive epistemology of the self. For, as we are now in a position to appreciate, a capacity for reflective self-cognition is a constitutive feature of rational autonomy. Again, this is because an autonomous activity is one that the subject can regard herself as actively engaged in—as in the relevant sense *doing*—and therefore one that has pure apperceptive form. From this it follows that the presumption of rational autonomy—a presumption, we saw, that Kant regards as a necessary feature of reason *as such*—is necessarily accompanied by the presumption of a capacity for reflective self-cognition. This has two consequences, both of which bear on the question of whether Kant is entitled to this presumption. First, the critique of pure reason cannot coherently suspend judgment as to whether we are "really" actively applying in thought the principle we take ourselves to be, for this would amount to a performative contradiction: critique is a reflective procedure *in conception*, so to engage in critique is to presume that we are conscious of the nature of our own thinking activity. Second, if Kant is correct that reason presumes its own normative authority, the critique of pure reason and *its attendant presumption of a capacity for reflective self-cognition* express a constitutive demand of reason.[51]

At crucial junctures in the *Critique of Pure Reason*, Kant argues from substantive, *a priori* claims about the nature of any finite, discursive intellect. As transcendental claims about the possibility of theoretical cognition, such claims are neither empirically tested nor empirically testable.[52] But as I hope

[51] Of course, nothing I say here entails that Kant is entitled to the assumption of reason's autonomy, much less that we are capable of critique. My aim has been to show that Kant's presumption of a capacity for reflective self-cognition has its basis in a deeper normative commitment, not that he is entitled to that commitment.

[52] Cf. Strawson's famous dismissal of the Transcendental Deduction as "an essay in the imaginary subject of transcendental psychology" (*The Bounds of Sense*, p. 32). Strawson further objected that "we can gain no empirical knowledge of its truth" (ibid.).

to have shown, critique's reliance on purported items of self-cognition revealed through reflection evinces Kant's commitment to and conception of the normative autonomy of reason. This reveals the deeper significance of the methodological disagreement between Kant and our imagined philosophical naturalist. The naturalist posits empirical confirmation as a standard for philosophical inquiry, and as a consequence holds that psychological assertions of the sort that Kant invokes demand empirical backing. For Kant, however, empirical science is situated within, and ultimately validated as science by, a philosophical framework committed to the autonomy of reason, a commitment he regards as internal to reason. This conception of philosophy allows for the possibility of natural scientific inquiry into the self *regarded as an empirical object*. However, for the purposes of philosophical inquiry, empirical science cannot supplant the conception of reason as sovereign. As Kant says in the Introduction to the second *Critique*, "It is pure reason that itself contains the standard for the critical examination of every use of it" (KpV 5:15).

I hope to have given some indication of what this passage from the second *Critique* means. Its full significance will be revealed only after taking account of Kant's critique of *practical* reason, which I will begin to address in chapter three. There I will argue that the self-consciousness of moral agency is a form of pure apperception, and that this bears on whether a "deduction" of morality can be given. The material from the present chapter, however, will mostly bear on a different question, which I take up in chapter four: whether in the absence of such a "deduction" we are nonetheless entitled to regard ourselves as moral agents. The position Kant adopts in the second *Critique* is that our consciousness of the imperatival force of morality is a "fact of reason," and that this fact entitles us to regard ourselves as moral agents in lieu of even the possibility of a deduction of morality. I will argue that our status as moral agents is, for Kant, a fundamental item of practical reflection, and that the *Critique of Practical Reason* likewise presumes a capacity for reflective self-cognition grounded in the autonomy of reason.

CHAPTER THREE

Kant's Fact of Reason as Pure Practical Apperception: Why a Deduction of the Moral Law is Impossible

§3.0 In his earlier moral writings, Immanuel Kant made several attempts to prove, from non-moral premises, that human beings are morally obligated. The last and best-known of these attempts takes place in the third and final chapter of the *Groundwork of the Metaphysics of Morals* (1785).[1] There Kant sets out to provide a "deduction" of the moral law, a kind of argument that, if successful, would establish that the moral law applies to all possible rational agents, and in particular would establish that morality applies to human agents as a system of categorical imperatives.[2] Such an argument aims to validate human agents' self-conception as moral agents, that is, their self-conception as free agents capable of acting (or abstaining from some action) entirely on the basis of their discursive representation of some candidate action as morally required (or morally prohibited). It would for that reason seem crucial to this argument that it proceed entirely from non-moral premises, including any suppositions about the legitimacy or veridicality of moral consciousness. Indeed, Kant openly worries about a "kind of circle" contained in the opening statement of the deduction, and he devotes much of the rest of the chapter to overcoming the circle he previously identified (G 4:450).

However, with the subsequent publication of the *Critique of Practical Reason* (1788), Kant disavows any attempt to provide a deduction of the moral law, asserting that such an argument is both unnecessary and—what will be the ultimate conclusion of this chapter—*impossible*. In its place Kant insists that our recognition of the unconditional authority of the moral law is grounded in a "fact of reason."

---

[1] Kant attempted several deductions of morality during his critical period, the last of which was given in *Groundwork* III. Cf. Henrich ("The Concept of Moral Insight and Kant's Doctrine of the Fact of Reason").

[2] In beings such as ourselves, who desire ends that sometimes conflict with what the moral law requires, morality is felt as an obligation or constraint. Kant's use of the term 'imperative' is meant to capture this feature of morality (G 4:414).

Kant's invocation of an unargued-for "fact" that grounds human morality has seemed to many commentators to be a retreat to precisely the sort of pre-critical dogmatism he criticizes in the *Critique of Pure Reason*. Karl Ameriks alleges, for example, that with the doctrine of the fact of reason, Kant "frankly acknowledged that his practical philosophy was 'dogmatic' and that only his theoretical philosophy was to be called Critical."[3] Such an interpretation is underscored by Kant's insistence that freedom of the will is a necessary condition of moral obligation,[4] which by the lights of the first *Critique* entails that moral obligation cannot be an item of theoretical knowledge. As a consequence, critical commentary has tended to focus on whether the later doctrine is in fact dogmatic, as well as whether (or to what extent) this doctrine represents a genuine shift in Kant's thinking.[5] In addition, commentators have tended to frame their discussions of the fact of reason in terms of a contrast with the deduction Kant attempts in the *Groundwork*. As a consequence, critical commentary has focused almost entirely on whether a deduction of morality is required for Kant to avoid dogmatism. Among Kant interpreters, the question concerning why Kant thought that a deduction of the moral law is *impossible* hasn't gone unanswered so much as not even asked.[6]

---

[3] "Kant's Deduction of Freedom and Morality," p. 72.

[4] Cf. G 4:447; KrV Bxxviii-xxix; KpV 5:4n, 5:28-29, 5:42ff.

[5] At one end of the spectrum is Karl Ameriks, who deems the doctrine of the fact of reason a "great reversal" in Kant's moral philosophy (*Kant's Theory of Mind*, p. 226). Likewise, Henry Allison claims that the *Groundwork* "does contain a serious attempt to provide a deduction (from nonmoral premises) of the moral law and/or the categorical imperative," which cannot be said of the doctrine of the fact of reason (*Kant's Theory of Freedom*, p. 280 fn. 2). Dieter Henrich claims that Kant is ambiguous on this point, and that one reason for this ambiguity is Kant's failure to distinguish stronger and weaker forms of a deduction; ultimately, according to Henrich, the weak deduction of morality in *Groundwork* III contains an implicit concession that "the consciousness of freedom has a presupposition in moral consciousness" ("The Deduction of the Moral Law: The Reasons for the Obscurity of the Final Section of Kant's *Groundwork of the Metaphysics of Morals*," p. 338). Jens Timmerman further argues that normative considerations were meant to "confirm" the results of the deduction in *Groundwork* III, and that we should therefore not assume that "Kant thought we had direct evidence of freedom independent of the recognition of normative laws" ("Reversal or Retreat? Kant's Deductions of Freedom and Morality," p. 82). At the other end of the spectrum is H. J. Paton, who claims that the *Groundwork* itself denies the possibility of a justification of morality from non-moral premises (*The Categorical Imperative: A Study in Kant's Moral Philosophy*, pp. 255-256).

[6] There are notable exceptions. Dieter Henrich accounts for the impossibility of what he calls a "strong" deduction of the moral law by invoking Kant's claim that "freedom has a presupposition in moral consciousness," i.e., that we come to regard ourselves as free only because we think of ourselves as morally obligated ("The Deduction of the Moral Law: The Reasons for the Obscurity of the Final Section of Kant's *Groundwork of the Metaphysics of Morals*," p. 338). The relevance of

I hope to redress this oversight. My aim in this chapter is to explain and defend Kant's claim that a deduction of the moral law is impossible. In doing so, I will leave largely unaddressed the attempted deduction of the *Groundwork* and the specific shortcomings of that argument. I postpone extensive discussion of that argument to chapter four. By reversing the usual order of exposition, I aim to elucidate Kant's fact of reason doctrine on its own terms, that is, independently of its relation to the prior argument. Furthermore, by setting forth in this chapter what I take Kant's fact of reason to be, I can then argue, in the next chapter, that Kant's *Groundwork* position already includes the conception of moral consciousness articulated more fully in the second *Critique*. (As we will see, Kant's views on the nature of moral consciousness shape the deduction that he attempts in the *Groundwork*, but the relevance of such views to that attempt are better understood in light of Kant's later position.) I will also set aside the question of whether the position Kant ultimately defends in the second *Critique* is "dogmatic" in Kant's sense, again postponing that discussion to the next chapter. Instead, the aim of this chapter will be to articulate and defend an interpretation of Kant's "fact of reason," and to show the relevance of this interpretation to Kant's claim that a deduction of the moral law is impossible.

According to the interpretation offered here, the fact of reason is the agent's pure apperception (in a sense to be explained) of her autonomous agency. I will argue that the pure apperceptive character of autonomous agency entails that such consciousness cannot form the basis of a deduction of the moral law. But before proceeding to my interpretation, some stage setting is required. In §3.1, I provide additional background to Kant's shift to the fact of reason doctrine, in order to provide necessary context for the discussion and state more precisely what Kant means by a "deduction." In

---

this claim to the possibility of a deduction of the moral law will become apparent in later sections, but for now notice that this claim pushes the problem back one step: we can now ask *why* our sense of freedom presupposes "moral consciousness." Less sympathetic commentators may be tempted to think that Kant renounced the possibility of a deduction only because of his own repeated failed attempts. Such a view is explicitly put forth by Ameriks ("Kant's Deduction of Freedom and Morality," p. 67).

§3.2, I defend, on both philosophical and textual grounds, my account of the fact of reason, and anticipate and respond to a possible objection to my account. In §3.3, I explain why, in light of this account as well as other considerations, a deduction of the moral law is impossible.

§3.1 In a well-known comment in the *Critique of Practical Reason*, Kant asserts that our "consciousness of [the] fundamental law" of pure practical reason is a "fact of reason" [*Faktum der Vernunft*] (KpV 5:31). This "fundamental law," which he elsewhere labels the 'moral law' (5:32), finds expression in the Formula of Universal Law. The latter sets forth the familiar Kantian standard of universalizability, the requirement to act only on those principles whose universal adoption one could concomitantly legislate (5:30).[7] The moral law is the fundamental requirement of morality, and it is through the fact of reason that the moral law "announces itself as originally lawgiving" (ibid.).

In the second *Critique*, Kant makes several allusions to the so-called "fact of reason." Unfortunately, Kant's various allusions and descriptions are hardly equivalent, and it is far from obvious how exactly these claims relate to each other.[8] To wit,

(i) At the beginning of the Preface to the second *Critique*, Kant asserts that pure practical reason "proves its reality and that of its concepts by what it does [*durch die Tat*]" (KpV 5:3). In chapter four, we will see that Kant's reference to the *Tat* (deed) of pure practical reason is an allusion to the fact of reason.[9]

[7] In the *Groundwork*, Kant provides three distinct formulations of the moral law, claiming that each is a different expression of one common principle of pure practical reason. (See footnote 2 of the introduction to this dissertation.) Here I will be focusing exclusively on the Formula of Universal Law. In the *Critique of Practical Reason*, Kant gives pride of place to the Formula of the Universal Law as the principle of pure practical reason. I believe this is because that formulation best captures what Kant takes to be distinctive about pure as opposed to empirical practical reason: pure practical reason is a capacity to transcend our sensible nature by adopting maxims on the basis of a purely formal consideration, namely, a maxim's lawgiving form.

[8] In addition to the list provided, see also KpV 5:91 and 5:104.

[9] That is not to say that Kant's notion of a *Tat* (deed) of reason is equivalent to his notion of a *Faktum* (fact) of reason, as some scholars have claimed. According to the view presented here, the "fact" of our moral consciousness consists in our

(ii) Later in the Preface, Kant claims that practical reason "furnishes reality" to the concept of *freedom* "by means of a fact" (5:6, my emphasis). There Kant asserts that freedom is a "supersensible object of the category of causality" and thus something of which we, as discursive subjects who depend on sensibility for theoretical cognition, could not have theoretical knowledge (ibid.).

(iii) In Chapter One, Kant officially introduces the "fact of reason" as "consciousness of [the] fundamental law [of pure practical reason]" (5:31, my emphasis). By means of such consciousness, the moral law "forces itself upon us of itself as a synthetic a priori proposition that is not based on any intuition" (ibid.). Moreover, what Kant refers to as *the* fact of reason is, according to him, the "sole" fact of reason, through which the moral law "announces itself as originally lawgiving" (ibid.).

(iv) Later in that same chapter, Kant refers to a "fact in which pure reason in us proves itself actually practical," which he identifies as "autonomy in the principle of morality by which reason determines the will to deeds" (5:42, my emphasis). Kant claims that this fact "is inseparably connected with, and indeed identical with, consciousness of freedom of the will" (ibid., my emphasis).

(v) Kant also seems to suggest that the fact that we are morally obligated is the relevant fact, in his claim that "the moral law is given…as a fact of pure reason of which we are a priori conscious and which is apodictically certain" (5:47).

(vi) In the final section of Chapter One, Kant appears to assert that the fact of reason is the will's determination by the moral law: "[t]he objective reality of a pure will or, what

pure apperception of the activity, or "deed," of pure reason in its practical operation. But since moral consciousness, as a mode of pure apperception, is constitutive of the very activity of which it is a consciousness, nothing substantive hangs on a terminological distinction between them. We can, for the purposes of conceptual taxonomy and terminological consistency, regiment the notions of *Tat-* and *Faktum der Vernunft* so as to make this distinction, but in doing so we ought to keep in mind Kant's own terminological imprecision in regard to these very issues.

> is the same thing, of a pure practical reason is given a priori in the moral law – for so we may call a determination of the will that is unavoidable even though it does not rest upon empirical principles" (5:55).

Despite the terminological imprecision that plagues Kant's many references to the fact of reason, I take the following to be sufficiently evident: (i) Kant believes that we possess some sort of consciousness of the imperatival force of morality; (ii) this consciousness is not derived from some antecedent fact or data of which we are aware—it is not something we base upon a prior consciousness of the will's freedom or some given intuition; and finally (iii) this consciousness provides some sort of entitlement to regard ourselves as free beings who face moral obligations. Thus, let us say for now that the fact of reason is our fundamental *consciousness* of what morality requires, and that this consciousness includes our recognition that such requirements make unconditional demands on us.[10] The nature of this "consciousness" will be clarified and explained as we proceed.

It may seem arbitrary or question-begging to stipulate at the outset that the fact of reason is our *consciousness* of the unconditional authority of morality. In light of other available interpretations, why not claim instead that the fact of reason just *is* the fact of our moral obligation, or the will's determination by the moral law (or—what for Kant amounts to the same—the exercise of the will's autonomy)? I believe that Kant's vacillation between describing the relevant "fact" as consciousness of moral obligation, and describing it as the very fact *of* moral obligation, reflects his conviction in the second *Critique* that moral consciousness provides the sole possible ground for regarding ourselves as morally obligated. It is especially noteworthy that in several of the passages cited above, Kant is concerned with establishing the "reality" of various *a priori* concepts, such as freedom and the moral

[10] This paragraph ignores one complication: the fundamental law of pure practical reason is only felt as a demand or "imperative" in beings subject to non-rational (i.e., sensible) "needs and motives" (KpV 5:32). This would not be the case for a "holy will," one that "would not be capable of any maxim conflicting with the moral law" (ibid.). Cf. G 4:412-413 and 4:439. See also footnote 2.

law. For this reason, there is a connection between the fact of moral consciousness and the fact, assuming it is a fact, that we are morally obligated, for Kant insists that the former is our only rational basis for thinking that we are so obligated. Moreover, if, as I shall argue, moral consciousness is the agent's pure apperception of the activity of moral deliberation and action, it shouldn't surprise us that Kant likewise vacillates between regarding the relevant fact as the "consciousness" (i.e., as we shall see, the pure apperception) of this activity and the activity itself (the will's "determination" according to moral dictates). In any case, let us proceed according to the stipulation made above. Its ultimate vindication will consist in its explanatory role in a thorough and systematic interpretation of the text.

As mentioned above, Kant invokes the doctrine of the fact of reason in place of any attempt to give a "deduction" of the moral law, which he now claims is neither necessary nor possible (5:42ff.). We saw that this stance is an apparent reversal of the position he espoused three years earlier in the *Groundwork of the Metaphysics of Morals*. The first two chapters of the *Groundwork* follow the "analytic" method, and are concerned to show (among other things) what the content of morality would be, *if* there were such a thing. Through analysis of the concept of a categorical imperative, Kant derives a version of the Formula of Universal Law (G 4:420-1). Later, at the beginning of the third and final chapter ('*Groundwork* III'), Kant argues that this principle is the principle of freedom, in the sense that the principle expresses the constitutive standard of autonomous volition; on this basis he concludes that "a free will and a will under moral laws are one and the same" (4:447). But given the limitations of our knowledge, we human beings cannot establish, and are not justified in assuming, that we are free in the absolute or "transcendental" sense required for moral obligation.[11] Thus, at this stage of

[11] Although Kant never uses the qualifying term 'transcendental' in the *Groundwork*, his discussion of freedom as having a positive and negative component corresponds to the way he describes transcendental freedom elsewhere. In the second *Critique* Kant defines transcendental freedom, negatively, as "independence from everything empirical and so from nature generally" (KpV 5:97), or the causal independence of the agent's practical choices from her natural impulses. In the first *Critique* Kant defines transcendental freedom, positively, as "the faculty of beginning a state from itself" (KrV A533/B561). In the context of Kant's practical philosophy, reason's power to causally effect a state absolutely "from itself," that is, without relying at all on sensible inclination as a source of motivation, is autonomy. The Formula of Universal Law is the constitutive principle of the will's autonomy, in the sense that it supplies the standard for what autonomous action is.

the argument Kant does not take himself to have shown that moral principles apply to us as categorical imperatives, only that they would if we were free. Even granting that Kant's analytic method sets forth the source and content of morality *assuming there is such a thing*, Kant has not thereby established that anyone is in fact morally obligated. Indeed, for all that he takes himself to have shown he still voices the worry that morality might be a "phantom of the human imagination" (4:407) or a "chimerical idea without any truth" (4:445).

There are two things to note about this worry. First, Kant himself does not seem to harbor any serious doubt that we are morally obligated. The discussion of *Groundwork* III clearly demonstrates Kant's conviction that we "take an interest" in morality (4:448ff.), where what he means by this is that we are capable of acting morally for its own sake (4:413n). Kant held that our capacity for moral motivation is our capacity to act on the basis of principles of pure practical reason. Thus, when Kant asserts that we are able to "take an interest" in morality, he is referring to our putative capacity to act on the basis of *pure* (as opposed to empirical) practical reason—our power to act on the basis of principles of practical reason whose motivational efficacy is absolutely independent of our various empirical desires, incentives, and inclinations. But despite Kant's conviction that we possess this capacity, in the *Groundwork* he sets for himself the critical task of explaining the possibility of our interest in morality—hence the possibility of pure practical reason—so as to certify the legitimacy of morality on the basis of considerations other than those of morality itself. Such is the project of a "deduction" of the moral law, which he undertakes in the bulk of the remainder of *Groundwork* III.[12]

This brings us to the second point. In the *Groundwork*, Kant explicitly claims that providing a justification of morality belongs to the broader and more systematic aims of reason's "critique"—in

[12] That Kant believed all rational beings take an interest in morality is also evident from his remarks at the conclusion of the deduction, where he claims that "common human reason confirms" its results: "There is no one – not even the most hardened scoundrel, if only he is otherwise accustomed to use reason – who, when one sets before him examples of honesty of purpose, of steadfastness in following good maxims, of sympathy and general benevolence (even combined with great sacrifices of advantage and comfort), does not wish that he might also be so disposed" (4:455).

this instance a critique of "pure practical reason" (4:391; 4:440; 4:445; 4:446). This is the sense in which answering the worry is a *critical* task. As with the critique of pure theoretical reason, part of its task is the establishment of synthetic, *a priori* truths. Recall a passage quoted in the previous chapter:

> That this practical rule [i.e., the Formula of Universal Law] is an imperative, that is, that the will of every rational being is necessarily bound to it as a condition, cannot be proved by mere analysis of the concepts to be found in it, because it is a synthetic proposition; one would have to go beyond cognition of objects to a critique of the subject, that is, of pure practical reason, since this synthetic proposition, which commands apodictically, must be capable of being cognized completely a priori. (4:440)

Kant will attempt the critique of pure practical reason in his third chapter, precisely through a "deduction" of the moral law. To appreciate the significance of this project, we must first take account of the broader meanings of these terms within Kant's philosophy.

As we saw in chapter two, the project of critique is most essentially the reflexive procedure whereby reason undertakes to validate some disputed claim concerning its own powers or identify the boundaries of its legitimate use. Kant hints at the reflexive character of critique in his claim, above, that establishing the authority of morality would require going "beyond cognition of objects to a critique *of the subject*" (4:440, my emphasis). The critical procedure, I said, both presupposes and exhibits the autonomy of reason, and I argued for a connection between this feature of critique and its being a *self-conscious* procedure. In particular, critique consists in reason's attempt to achieve self-cognition according to its own, self-consciously prescribed standards. Finally, I claimed that critique concerns itself with the "sources and limits" of some mode of rational cognition, but I postponed discussion of this aspect of critique to the present chapter. To fill in this gap in the account, let us now examine Kant's notion of a deduction.

While "critique" broadly consists in reason's self-examination, Kant reserves the term "deduction" for the specific kind of argument whereby some disputed claim about reason is legitimated or justified. Dieter Henrich has persuasively argued that Kant's notion of a philosophical

deduction was modeled on eighteenth-century legal deductions that resolved territorial disputes between German provinces. In Kant's time there existed the widespread practice of submitting "deductions" before a court in order to make claims on disputed territories. Rightful claims to land had a particular *origin*, and it was the function of a legal deduction to trace disputed claims to possession to the original fact or deed in consequence of which the disputant had possession of the territory. Analogously, the point of a philosophical deduction is to vindicate our "right" to apply an *a priori* concept, by tracing the use of that concept to the mental powers or activities from which it originates.

The structure of a deduction is displayed most clearly in the Transcendental Deduction of the first *Critique*. There Kant accounts for the objective validity of certain pure concepts of the understanding, or "categories," by showing that they are required for the possibility of self-conscious, empirical representation. So, for example, although no finite set of experiences could license a claim of causal connection among objects, or a claim concerning the relationship of substance to accident, Kant argues that it is only by discursively representing objects according to the categories of causality and substance that we can represent them *as objects* in our judgments about them, that is, *in thought*. (And so on for the other categories.) And as we saw in chapter one, each of these categories is itself associated with a distinct form of representational activity—a mode of the activity Kant calls "combination"—the awareness of which constitutes the pure apperception that Kant takes to be partly criterial of thought about empirical objects. Thus, the very cognitive powers that enable us to represent the empirical world in thought presuppose categorial modes of representation. From this it follows that the empirical world must actually conform to the categories if veridical judgments concerning the empirical world are to be possible. Setting the details of this argument aside,[13] I wish to emphasize only that Kant attempts to establish the possibility of synthetic, *a priori* knowledge by

[13] For ease of exposition, I ignore the complication that Kant provided two versions of the Transcendental Deduction, and refer instead to a single argument. This should not be understood as my taking a position on the question of what, if any, differences there are between them concerning their argumentative structure.

tracing this possibility to various more fundamental mental capacities he identifies as necessary preconditions of empirical judgment. In this way, the argument exhibits the general structure of a philosophical deduction. In Henrich's words, it "seeks to discover and examine the real origin of our [claim to knowledge] and with that the source of its legitimacy."[14]

In parallel fashion, a "deduction of the moral law" would establish the entitlement or "right" of rational agents such as ourselves to regard ourselves as *moral* agents, by grounding a capacity for moral motivation in some more fundamental rational capacity. In the passage quoted above, we see that the aim of a deduction of the moral law is included within the general ambitions of a "critique of the subject, that is, of pure practical reason" (4:440). The proposition that the moral law applies to us as an unconditional practical principle is, Kant tells us, a synthetic, *a priori* claim. As such, it cannot be justified on the basis of either conceptual analysis or experience and therefore, according to Kant, positively *requires* a critique of pure practical reason. I argued in chapter two that, necessarily, Kantian critique draws upon items of self-cognition made available through first-person reflection on our various rational capacities. In particular, I asserted that the Transcendental Deduction of the Categories makes reference to rational capacities, including our very status as self-conscious subjects, disclosed to us through reflection. We should therefore expect that the "more fundamental source" articulated in a deduction of the moral law would likewise be disclosed to us through reflection. However, in addition to this similarity, there is an important difference between the Transcendental Deduction and a putative deduction of the moral law. The former argument vindicates our right to apply categories to the empirical world, by showing that the objects of our perception actually stand in categorial relations. By contrast, in a deduction of the moral law it is sufficient to show that we are capable of deliberating and acting according to principles of pure practical reason—in short, that we *are* pure practical reasoners. Thus, it is not part of the task of a deduction of the moral law to show

[14] "Kant's Notion of a Deduction and the Methodological Background of the First *Critique*," p. 35.

that synthetic, *a priori* principles apply to given objects. In this respect, a deduction of the moral law is similar to the so-called "metaphysical" deduction of the categories, which established merely the discursive forms that our thought of objects *qua* objects must take, regardless of whether those forms apply to the objects themselves—indeed, regardless of whether objective cognition is possible for us at all. (I shall have much more to say about these issues and their significance for Kant's project in the next chapter.)

The generic form of a deduction is exhibited by the argument that Kant attempts in *Groundwork* III. There Kant attempts to ground the claim that, necessarily, the Formula of Universal Law is the fundamental practical principle of all rational beings, by tracing this claim to some necessary and universal feature of rational agency, one that entails the necessity of agents regarding themselves as morally obligated in their practical deliberation. Again, I will have much more to say about this argument in the next chapter, but very roughly, it goes as follows. First, recall that in the brief, "analytic" portion of *Groundwork* III, Kant argued that the Formula of Universal Law is a principle of freedom, in the sense that it sets forth the fundamental criterion or standard of autonomous volition.[15] Kant proceeds to claim that, necessarily, all rational agents act "under the idea of freedom," in the sense that we presuppose our own freedom whenever we deliberate about what to do. On this basis, Kant concludes we necessarily recognize as binding the principle of such freedom.[16] Since, according to Kant, the fundamental principle of freedom is the Formula of Universal Law, it follows that as

[15] For Kant's discussion of the "analytic" versus "synthetic" methods in the *Groundwork*, see G 4:392. The terms 'analytic' and 'synthetic' apply to distinct methods of philosophical argumentation, and should not be confused with the analytic-synthetic distinction among judgments. Put generally, the "synthetic" method seeks to establish the objective validity of some body of knowledge, by establishing the principles or capacities that ground such knowledge, and showing how they do so. By contrast, the "analytic" method takes for granted some body of knowledge and regresses to the conditions of the possibility of such knowledge. Cf. Prol 4:263-4 and 4:274-5.

[16] I said earlier that, because of a restriction on our capacity for knowledge, we cannot be justified in the belief that we are transcendentally free. Given that restriction, we might wonder what entitles us to represent our actions "under the idea of freedom." Kant's answer is that we are only entitled to regard ourselves as free "in a practical respect" (G 4:448). In other words, Kant does not claim that our freedom is an item of *theoretical* knowledge, only that the assumption of our freedom is valid for the practical use of reason.

rational beings we recognize as binding the Formula of Universal Law and the particular categorical imperatives that issue from it.

Now, if it were the case that we only regarded ourselves as free because we took ourselves to be morally obligated, then the presupposition of freedom would itself presuppose our recognition of the unconditional authority of the moral law. And if this were the case, Kant's aim of providing a "deduction" of the moral law—that is, a vindication of our "right" to apply moral concepts to our action, on the basis of the cognitive activities from which these concepts originate—would plainly be threatened. Indeed, in the chapter's ensuing discussion Kant raises and tries to disarm precisely that concern. In particular, he raises the concern that we only "take ourselves to be free…in order to think ourselves under moral laws," and hence that there is a "a kind of circle" involved in our attempting to legitimate morality on the basis of the fact (supposing it is a fact) that rational agents act under the idea of freedom (G 4:45).

With this as background, I am now able to state the apparent difference in strategy between the *Groundwork* and second *Critique* more precisely. While in the *Groundwork* Kant argues from the presupposition of our own freedom to the authority of moral principles, in the second *Critique* Kant argues that our consciousness of the unconditional authority of the moral law is a "fact of reason," and that it is only on the basis of this "fact" that we are aware of our freedom. It would therefore seem that Kant abandons any attempt at a "deduction" of morality from freedom precisely because he holds that it is only by representing the moral law as authoritative that we come to regard ourselves as free. As he puts it in the Preface to the second *Critique*, the moral law is the "*ratio cognoscendi*" of freedom (KpV 5:4n).

This goes some way toward explaining why Kant came to believe that a deduction of the moral law is impossible. For if Kant later affirmed that our recognition of moral constraints is the ground of our regarding ourselves as free, he must have recognized that at least one strategy for deducing

morality is blocked. If the representation of moral requirements is the sole ground of our consciousness of freedom, then we cannot appeal to our self-conception as free beings to justify the claim that we are morally obligated. Kant appears to have thought that no other strategy is available. Nonetheless, it remains unclear *why* Kant thought that our consciousness of freedom is ultimately rooted in our recognition of moral requirements.[17] To see why, we must take account of Kant's views on moral consciousness, which I take up in the next section.

§3.2 In chapter one, I explicated Kant's notion of "pure apperception," which I claimed was the distinctive mode of self-consciousness in which one is conscious of oneself not as an object of representation but rather as a thinking subject, the 'I' that is engaged in the activity of thinking. On the basis of textual evidence from the A and B Transcendental Deductions, I characterized pure apperception as a kind of "agential" consciousness:

> Consciousness of myself *as subject* consists in my consciousness of engaging in the activity of combination, where (i) this consciousness is constitutive of that very activity, and (ii) this consciousness constitutively identifies me as the "agent" of that activity.

My aim in this section is to argue that, on Kant's view, moral consciousness likewise exhibits the form of pure apperception. In other words, the form of pure apperception is exhibited not only in one's consciousness of oneself as a thinking subject, but moreover in one's consciousness of oneself as an autonomous agent who stands under moral obligations.

I have just now identified "moral consciousness" with "consciousness of oneself as an autonomous agent who stands under moral obligations." A natural worry is that in doing so, I have

[17] See footnote 6.

elided the distinction between (i) consciousness of myself as an autonomous agent and (ii) my consciousness of the content and imperatival force of morality—the "fact of reason" as it is exhibited in my practical thought. Indeed, the conclusion of this section will be that, according to Kant, (i) and (ii) amount to fundamentally the same thing: my consciousness of myself as an autonomous agent just is my pure apperception of deliberating and acting on the basis of my recognition of moral demands. Furthermore, my consciousness of the content and imperatival force of morality is a *practical* consciousness, consisting in my recognition of moral demands in practical thought and action. When we combine these claims, we see that (i) and (ii) differ only in the descriptive emphasis they put on what is really a single, distinctive kind of activity, namely, moral deliberation and action. The first description emphasizes the pure apperception that is constitutive of this activity, the consciousness of *myself* as the agent of (moral) deliberation and action. The second description emphasizes the moral imperatives that structure and define my thought and action. But the activity they describe is the same. However, in order to secure this conclusion, let us for the time being make a nominal distinction between them.

Whether they are in fact distinct will make a difference to the kind of "deduction" of morality we might attempt. If an agent's recognition of the unconditional authority of the moral law is a constitutive feature of her consciousness of her autonomy, then we cannot appeal to the latter in an argument attempting to establish her "right" to regard herself as the type of agent who recognizes and is motivated by moral considerations. By extension, agents' capacity for this form of self-consciousness cannot supply the basis for a "deduction" establishing that the moral law applies to them.

Before proceeding to my argument, I need to say a bit more about autonomy and the relationship of this notion to freedom and morality, as well as provide a working definition of "consciousness" of this capacity. In chapter two, I claimed that pure reason in its self-critical operation

exhibits a kind of "autonomy," in the sense that the normative standards structuring its activity are reason's own. On this view, autonomy is regarded as the property that rational capacities possess insofar as they are *self*-governing. In the example from chapter two, pure reason engages in self-critique according to the standards that issue from pure reason itself; for this reason, I said, the achievement of the critical goal does not consist in meeting an extrinsic normative demand. But as I admitted in the last chapter, my use of the term 'autonomy' is somewhat broader than Kant's, who restricts its usage to describe the capacity for pure *practical* reason, or the capacity of the rational will to determine itself only according to its own, pure principles. The rational will is a species of the more general "faculty of desire" [*Begehrungsvermögen*], "the faculty to be, by means of one's representations, the cause of the objects of these representations" (MM 6:211). In non-rational beings, that faculty is determined entirely by what Kant calls "sensible impulses" (6:213-4). But *our* actions are not dictated by our impulses; as rational agents we are able to reflect on the possible grounds of our actions and determine whether they are, in some sense, good reasons for action. We possess a *will*, which Kant identifies with practical reason (MM 6:213; G 4:412). "Will" or practical reason is thus the distinctive form that the faculty of desire takes in rational agents, an ability to represent actions in the form of principles (or "maxims") and, on the basis of one's determination that some action would be good to do, choose to perform that action (G 4:412).[18]

Kant argues in the *Groundwork* that, necessarily, any rational will is an *autonomous* will. (That is his *Groundwork* position at least; in a later work he seems to retract this claim.[19]) While the will is

[18] Cf. Korsgaard ("Acting for a Reason," pp. 208-214). In the *Groundwork*, Kant makes the stronger statement that the will is "the capacity to act *in accordance with the representation* of laws" (G 4:412). Following Reath, I take Kant to mean that the will is a capacity to act according to the representation of *objective practical principles*, that is, on the basis of reasons that the agent judges as valid reasons for action for all rational agents ("The Categorical Imperative and Kant's Conception of Practical Rationality," pp. 77, 89 fn. 24). Such principles would include both hypothetical and categorical imperatives. An objective practical principle properly becomes law only when it is unconditionally binding on all rational agents, as categorical imperatives are. Kant sometimes uses 'law' in the weaker sense of 'principle'. For instance, he defines a maxim as "a subjective law by which we actually do act" (LE 27:1427).

[19] Rel 6:26n.

defined as practical reason, autonomy in this more restrictive sense is the capacity to act on the basis of principles of *pure* practical reason—an ability to be motivated by pure reason alone, without any motivational "assistance" issuing from sensibility. We can think of an autonomous being as one for whom, for example, the representation that some course of action will maximize her long-term happiness does not settle the question of whether she ought to pursue it. From the standpoint of pure rational deliberation, happiness, considered as a practical end, is itself an object of critical assessment. As I stated in §3.1, the fundamental principle of pure practical reason is Kant's Formula of Universal Law. Autonomy, then, is the capacity to act (at least partly) on the basis of one's representation that some maxim is suitable for universal adoption, as determined by this principle.[20] For finite rational agents, autonomy is the capacity for moral motivation.[21] Such a capacity is what Kant refers to in the *Groundwork* as our ability to "take an interest" in the moral law (G 4:449).

The positive characterization of autonomy in terms of a specifically moral capacity supplies a substantive conception of the will's freedom. I will have more to say about this in the next chapter, but for now it suffices to say that for Kant, 'freedom' has both a negative and positive significance (G 4:446).[22] Understood negatively, the will's transcendental freedom is the will's freedom *from* natural-causal determination, or "that property of [the will] that it can be efficient independently of alien causes *determining* it" (ibid.). But the Formula of Universal Law, as a constitutive principle of the will's autonomy, supplies a contentful standard for what a negatively free will *would* choose, if it were to

[20] I say 'at least partly' because merely permissible maxims are such that both they and their contraries have universal form, so whether to adopt some merely permissible maxim is not settled by the requirement to act only on universalizable principles.

[21] Here I restrict the term 'moral motivation' to exclude agents who by their very nature never violate the Formula of Universal Law and thus for whom the "requirement" of universalizability is not felt as a demand.

[22] See footnote 11.

freely determine itself according to its own, rationally-prescribed standards.[23] The Formula of Universal Law is thus a "principle of freedom" in the sense that it is the internal standard of the will's autonomy, i.e., its *positive* freedom.[24] In light of this distinction, we should recast the main question of this chapter as follows: is it possible that our consciousness of our *autonomy*—our ability not simply to exercise externally-causally-undetermined choice (negative freedom), but additionally to exercise that capacity purely on the basis of considerations that emanate from our own pure reason (positive freedom)—could serve as the basis of a deduction of the moral law?[25]

In order to give sense to this question, let me say a brief word about what I mean by "consciousness" of autonomy. In the Transcendental Deduction, Kant is concerned to show (among other things) what is involved in the kind of self-consciousness subjects possess insofar as they are conscious of themselves *as subjects*. Similarly, we can ask what, on Kant's view, is involved in the kind of self-consciousness autonomous agents might possess insofar as they are conscious of themselves *as autonomous agents*—insofar as they possess a kind of self-consciousness that constitutively identifies them as deliberating and acting according to the standards of their own *pure* reason, and thus as exhibiting the capacity for positive freedom.

---

[23] In the *Groundwork*, Kant claims that the merely negative characterization of freedom "is unfruitful for insight into its essence," but that a positive conception of freedom "flows from" the negative one (4:446). A positively free agent is not causally necessitated by natural laws (i.e., is negatively free), and moreover can be motivated to act on the basis of pure reason alone, without requiring any motivational "assistance" from sensible desires. Autonomy, understood here as the capacity to act on the basis of principles of pure practical reason, is thus identical with positive freedom. Kant maintains the distinction between positive and negative freedom in the second *Critique* (5:29).

[24] This is explicitly affirmed by Kant in the *Groundwork* (G 4:446-7) and entailed by the second *Critique's* discussion of the relationship of freedom to the moral law's requirement to act only on those maxims that possess "lawgiving form" (KpV 5:28-30). In the latter discussion, Kant asserts, first, that only a transcendentally free will can determine itself according to the mere lawgiving form of its maxims; and second, that the "lawgiving form" of a maxim is the sole "determining ground" possible for a transcendentally free will. He concludes in the ensuing remark that "freedom and unconditional practical law reciprocally imply each other" (5:29).

[25] Note that possession of the capacity for positive freedom entails possession of the capacity for negative freedom. As I argue in chapter four, the converse does not hold.

I claim that, on Kant's view, such self-consciousness consists in the pure apperception of moral deliberation and action, in a sense to be explained. My argument for this thesis consists of two parts.

First, I will show that difficulties arise for the possibility of such self-consciousness that parallel the difficulties discussed in chapter one for the possibility of consciousness of oneself *as subject*. We can imagine an argument, roughly parallel to Hume's skeptical argument against self-consciousness, that calls into question the possibility of being conscious of oneself as an autonomous agent. As a first approximation, that argument would say that we are empirically conscious of nothing that could plausibly be construed as our autonomous selves. To secure an even stronger conclusion, we might argue as follows: since (P1) the concept of autonomy is one whose instantiation cannot be given empirically;[26] and since (P2) the only way a subject could be conscious of anything is to be conscious of it empirically; it would follow that (C) consciousness of autonomy is impossible. *A fortiori* consciousness of oneself as an autonomous agent is impossible. Stated as such, this argument does not give the full measure of the difficulty of accounting for the relevant form of self-consciousness. For it is possible to deny the claim that empirical consciousness exhausts our capacity for representation. We must state the problem at the same level of generality we did in addressing consciousness of oneself *as subject*: the mode of presentation involved in the self-consciousness of autonomous agency cannot be in the manner of representational content "given" to consciousness. In what follows I will argue for the necessity of this claim. Given the parallel with the "problem of self-consciousness" advanced in chapter one, this suggests that Kant should have construed the self-consciousness of autonomous agency on the model of pure apperception.

[26] This follows from Kant's characterization of transcendental freedom as "independence from everything empirical and so from nature generally" (KpV 5:97). Objects of experience stand in causal relations to each other according to natural laws, but the concept of a free will is the concept of an uncaused causality, hence one that isn't subject to the causal nexus of nature. Thus a free will cannot be an object of empirical cognition.

Second, I will provide textual evidence for thinking that Kant does indeed conceive of such self-consciousness on the pure apperceptive model. We shall see that one of the virtues of my interpretation is that it is able make sense of some of Kant's more opaque remarks concerning the nature of moral consciousness.

To begin, note that in order for autonomy to constitute a genuine form of rational self-government, exercises of this capacity must be such as to be *essentially* available to self-conscious reflection. To be sure, not every exercise of pure practical reason must be attended by an explicit, conscious judgment that *I* am doing something. We see this most clearly when we consider merely permissible actions: sometimes I just eat a maple syrup donut without giving it much (or any) explicit thought, though the consideration that this is a permissible thing to do still regulates what I am doing at some level. And similarly when I do what is required, or abstain from doing what is impermissible: I might stare longingly at *your* maple syrup donut, but I don't steal it from you when you are looking away, and this is because I judge that such an action would be wrong (i.e., based on a maxim that is not universalizable).[27] In such a case, I abstain from performing some action X because I judge it to be wrong, but this does not entail that my abstaining is accompanied by an explicit judgment of the form "I am not doing X because doing so would be wrong." However, in order for these actions to constitute genuine exercises of autonomy, it cannot be the case that it would be impossible for me to discursively represent what I am doing (as expressed in the maxim of my action) or that *I* am the one doing it. If it were, then I would stand to my own actions as a passive observer does to some bodily movements, and I could not be seen as exercising a capacity for choice [*Willkür*]. In the *Metaphysics of Morals* Kant defines 'choice' precisely as a kind of determination to act joined with "consciousness of

[27] For ease of exposition, I speak in this chapter as if rational human beings are autonomous in Kant's sense. However, nothing important turns on this convention. Since I am interested only in the conditions and implications of autonomous agency, one can read these sections simply as specifying what necessarily follows from the supposition that we are such agents.

the ability to bring about its object by one's action" (MM 6:213). Thus the capacity to choose is, in this sense, an essentially self-conscious capacity.[28] Moreover, to represent myself as performing an action is to represent myself as doing something *for a reason*, so I must be able to consciously represent my reason for action. From this it follows that autonomous agents, i.e. those who are capable of acting on the basis of the universalizability of their maxims, must be able to represent to themselves the suitability of their maxims for universal adoption. I do not take this to mean that they must have a sophisticated grasp of the letter of Kant's universality requirement, just that they be able to represent in some way that what they are doing is what anyone similarly situated is required or permitted to do.[29] Moreover, in representing their actions thusly, they at the same time represent *themselves* as capable of being motivated by a consideration other than their natural inclination toward some object or even their own happiness; to that extent, they represent themselves as autonomous. So an exercise of autonomy is by its very nature such as to be possibly represented by the agent *as an exercise of autonomy.*

With this conclusion, we can see the first signs of a parallel between the self-consciousness of autonomous agency and the apperception characteristic of the representation of one's pure subjectivity. Recall the principle of the analytical unity of apperception:

> The [representation] **I think** must **be able** to accompany all my representations; for otherwise something would be represented in me which could not be thought at all, which is as much to say that the representation would either be impossible or else at least would be nothing for me. (KrV B131-2)

---

[28] Technically, what I have just defined as 'choice' [*Willkür*] is what Kant refers to in the *Metaphysics of Morals* as specifically *human* choice (MM 6:213). Kant defines the latter as the capacity to bring about some object on the basis of a *conceptual* representation of that object's desirability (which may be grounded in sensibility or pure reason), with an attendant consciousness of the ability to realize that object through one's action. This is the sense in which (human) choice is an essentially self-conscious capacity. But Kant also has a broader notion of choice that includes (non-human) *animal* choice. Unlike human choice, animal choice is determined by sensible impulses. It therefore does not involve the mediation of a conceptual representation, and thus is not a self-conscious capacity (at least not in the sense expressed by the first-person pronoun in explicitly self-conscious acts of will). I mention this technicality only to set it aside. Going forward all references to our capacity for choice will be to the specifically human choice described above.

[29] On a certain conception of belief, moral agents do not have to *believe* that what they are doing is universally required or permitted. An autonomous agent might very well be a thoroughgoing nihilist in the seminar room. The representation we are after is a distinctively practical one.

This principle sets forth a necessary condition on the possibility of a representation figuring in the content of a judgment: any representation that could serve as part of the content of a subject's judgment is such that it could be identified by the relevant subject as belonging *to* the subject, the selfsame "I" that *thinks* that content. And although Kant entertains the idea that a representation that could not be thought at all would be impossible, he does not wish to rule out representations that might be "in" a subject while at the same time being unavailable *to* that same subject in an act of self-conscious reflection. For that reason, he settles on the weaker claim that an unthinkable representation "at least would be nothing for me" (ibid.). We saw in chapter one that Kant was specifically concerned with the conditions under which *intuitions* can serve as the contents of a judgment; a consequence of the principle of the analytical unity of apperception is that "all manifold of intuition has a necessary relation to the **I think** in the same subject in which this manifold is to be encountered" (B132). Thus, an intuition's being "for me" consists in the possibility of its serving as the content of a judgment of the form "I think x."

The practical analog of an empirical intuition that is something "for me" is an empirical incentive on which I could elect to act by choosing a maxim that incorporates the end specified by that incentive into its content. (If this is a bit unclear, don't worry: I will have more to say about maxims and their significance for Kant's account of action momentarily.) But as we saw above, *all* action involves choice [*Willkür*], and choice is an essentially self-conscious capacity. Thus, while there may exist cognitively significant theoretical representations that are unavailable to self-consciousness, an action is by its nature such that I can explicitly represent myself as its agent. (This is not to claim that there is no practical analog of the representation that is unavailable to self-consciousness, for what we have said is compatible with the existence of, e.g., unconscious drives that regulate and determine behavior. The point is just that such behaviors do not involve choice.) Thus, the foregoing reflections imply the following principle:

> The representation '**I will**' must **be able** to accompany all my actions, for otherwise they would not be my actions.

That is, just as the possibility of the representation 'I think' is a condition of a representation's being available to self-conscious reflection, so the possibility of the representation 'I will' is a condition of an action's *being my action at all*. In an autonomous agent the representation 'I will' does not mark arbitrary choice (i.e., merely an exercise of negative freedom), but autonomous choice, the capacity to be motivated by pure practical thought. The formulation of our principle makes explicit that it must only be possible for an agent to represent 'I will'. But the fact that the agent is *necessarily* able to do so points to a mode of self-consciousness that essentially figures in autonomous agency: consciousness of oneself as an autonomous agent. Here again, an analogy is suggested: just as the analytical unity of apperception—the representation 'I think'—implies a self-consciousness fundamental to discursive subjectivity, which Kant in the first *Critique* calls 'pure apperception', so the representation 'I will' implies a self-consciousness fundamental to autonomous agency.

As with the self-consciousness fundamental to discursive subjectivity, the self-consciousness fundamental to autonomous agency cannot be construed in the manner of representational content "given" to consciousness. As a first pass, we can argue for this claim by supposing that in order for consciousness of one's autonomy to be given as content, it would have to be given as *experiential* content, and thus represented as belonging to the causal nexus of nature. Since representing oneself as autonomous involves representing one's actions as undetermined by natural laws—that is, includes the representation of oneself as negatively free—the concept of autonomy cannot be instantiated in experience, and so consciousness of oneself as autonomous cannot take the form of some experiential content.[30]

[30] See also footnote 26.

Correct as this argument is, it does not state the problem in its full generality, and in particular it does not address the possibility that consciousness of one's autonomy is the representation of a *non*-experiential content (for example, a supersensible intuition of an active soul). We see the generality of the problem when we attend to the fact that representing oneself as autonomous entails representing oneself as a locus of responsibility. The notion of responsibility I have in mind is metaphysical, not legal or moral. It does not by itself predicate of its object the suitability of any form of legal redress or any morally-laden responsibility to or for another person. Rather, in taking myself to be autonomous, I take myself to be the fundamental cause of my actions, thus the fundamental cause of any state of the world brought about through my actions. In the *Groundwork,* Kant describes this phenomenon as the taking up of a certain "standpoint," one where "by means of freedom we think of ourselves as causes efficient a priori" (G 4:450). Importantly, the standpoint of autonomy is not merely one of regarding oneself as an uncaused or unconditioned causality. Rather, to regard oneself as autonomous is to regard one's own reason as the fundamental "source" or "ground" of one's actions, and thus to regard oneself as having the power to begin a new state on the basis of one's own (pure) reasoning about the matter.[31] This is the metaphysical sense of 'responsibility' I have in mind.

If the consciousness of my autonomy were originally supplied in the form of some given content, then my fundamental representation of this capacity would be as a power that belongs to something distinct from my own reason. I would not have an unmediated representation of my actions as being absolutely "up to me." Rather, I would place the locus of responsibility in the object corresponding to this representational content, and whatever representation I would have of myself as autonomous would be mediated by the representation that *this object is "my autonomy.*" If this were the form of my agential self-consciousness, I could not make an autonomous choice; at most I could

[31] Cf. KrV A533/B561. In making this distinction, I again have in mind Kant's distinction between (mere) negative freedom and positive freedom, or autonomy.

represent a certain maxim as choice-worthy and confidently *predict* that "my autonomy" won't let me down. But choice [*Willkür*] is not prediction, for the latter does not involve any intention on my part (in Kant's idiom, any "determination of my will") to perform an action.[32]

This poses a problem for giving an account of the self-consciousness of autonomous agency, one that exactly parallels the problem discussed in chapter one. As I argued there, Kant was without question aware of this problem as it pertained to the possibility of the representation of our pure subjectivity, as expressed by "I think." Moreover, Kant certainly held that the representation of ourselves as autonomous could not be provided for in the manner of empirical content,[33] and it is unlikely that the generality of the problem would have been lost on him. So we should expect that the form of Kant's solution to this problem would be the same. Kant held that consciousness of myself *as subject* consists in the pure apperceptive awareness of rule-governed activities of combination. We should therefore expect Kant to hold that the consciousness of myself *as autonomous* is most fundamentally the pure apperceptive awareness of a rule-governed activity.

I submit therefore that the "fact of reason" is the *pure apperception* of pure practical deliberation and action—that is, the pure apperception of practical deliberation and action guided by the Formula of Universal Law. Accordingly,

---

[32] There is a less direct route to the same conclusion. Agency involves representing that the empirical world is a certain way, which requires that we be subjects of experience. Therefore, we are no less subjects of experience when we are agents. In particular, instrumental reasoning represents certain means as conducive to our ends *given the state of the empirical world*. In that respect, instrumental reason draws upon theoretical reason. In order to have a unified representation of oneself as *pursuing this end through these means because the world is in such-and-such a state*, one must represent the same 'I' as both subject and agent—and this entails that when we act, any condition on the representation of ourselves as subjects is likewise a condition on the representation of ourselves as agents. So just as the representation of the 'I' *as subject* cannot be in the form of a given representational content—something represented as an object of representation, and thus represented as alienated from the 'I' that "does" the representing—so the representation of the 'I' *as agent* cannot be in in the form of such a content. *A fortiori*, the representation of the 'I' as *autonomous* agent cannot be in the form of such a content.

[33] To cite just one example, Kant opposes the "standpoint" of autonomous agency to that whereby we represent ourselves "as effects that we see before our eyes" (G 4:450).

> Consciousness of myself *as an autonomous agent* consists in my consciousness of engaging in the activity of deliberating and acting according to the Formula of Universal Law, where (i) this consciousness is constitutive of that very activity, and (ii) this consciousness constitutively identifies me as the agent of that activity.[34]

Just as consciousness of myself as a subject is the pure apperceptive consciousness of the activity of combination, so consciousness of myself as an autonomous agent is the pure apperceptive consciousness of the activity of pure practical deliberation and action. In both cases, consciousness of the activity is not a higher-order representation, one that takes the relevant activity as its object but is distinct from it—again, that would be to construe the relevant mode of self-consciousness in the manner of a given representation with determinate content. Rather, the pure apperception of the activity of pure practical reason is constitutive of that very activity, and this is what makes it a form of *pure apperceptive* consciousness. Consciousness of myself *as an autonomous agent* is thus an essential and constitutive aspect of pure practical deliberation and action.

What, more precisely, is the relevant activity? Since the Formula of Universal Law is the fundamental principle of pure practical reason, it is to deliberate and act according to the standard that an agent's principle of action (her "maxim") should have the "form of a law," as specified by that principle. Kant asserts that we become "immediately conscious" of the moral law "as soon as we draw up maxims of the will for ourselves" (5:29). Let us consider this statement in two parts, beginning with the notion of an agent proposing maxims to herself. Typically, the maxims that the agent

[34] Here I dispense with the scare quotes around 'agent' in the term 'agent of that activity' that I used earlier. In the case of the representation of pure subjectivity, as expressed by "I think," the point of my calling it a form of agential consciousness was to highlight the fact that pure self-consciousness involved a consciousness of oneself as active, and thus constitutively identified the subject as the source—the "agent"—of that activity. The representation of oneself as an autonomous agent likewise includes the representation of oneself as the source of an activity, but moreover involves the representation of oneself as an agent in the usual, richer sense. That is, it also involves the representation of oneself as having the power to exercise choice [*Willkür*] in a determination of one's will, and therefore, according to Kant, as a kind of uncaused causality and potential causal ground of a new state of the natural world.

proposes define a range of circumstances C in which a particular act-type A is to be performed for purpose P.[35] Moreover, while maxims may be adopted because they are required by a universal or objective standard (and thus to that extent codify objective requirements), they are the agent's own, "subjective" principles of action, and are regarded by the agent "as holding only for [her] will" (KpV 5:19). Thus, properly articulated, an agent's maxim makes an ineliminable first-person reference to her own willing, and the generic form of a maxim is something like the following: *I will perform act* A *in circumstances* C *to achieve purpose* P.

Kant's reference to a process initiated by an agent "draw[ing] up maxims of the will for [herself]" implies that this process begins with the agent representing some maxim as potentially choice-worthy, i.e., as a candidate for adoption. Since the generic form of a maxim is specified in terms of what the agent (potentially) *wills*, this in turn suggests that practical deliberation is at every stage oriented *toward action*, even if some particular instance of practical deliberation fails to eventuate in action. This may seem like a trivial observation, but its significance can be underscored by contrasting such a conception of practical deliberation with a competing one. On Kant's account, practical deliberation is the activity of practical reason, which is the form that the faculty of desire takes in rational agents, i.e., those who are capable of choosing to act on the basis of their representation that some maxim is good or worthwhile. But to represent some maxim as potentially "good" in this sense is already to regard it as a potential object of choice [*Willkür*]. It is therefore not to regard it as possessing some abstract property ("goodness") that is taken to be practically relevant only at a later stage in the practical-deliberative process, as one might do in a theoretical representation of some aspect of the world the representation of which, taken by itself, is motivationally inert. Kant summarily

[35] Kant provides several examples of maxims in the *Groundwork* and second *Critique* (G 4:397-398, 4:422-423; KpV 5:30). I borrow the term 'act-type' from Korsgaard ("Acting for a Reason," p. 219) so as to distinguish the self-consciously willed behavior done in such-and-such circumstances for such-and-such purposes from the broader *action*, which is individuated by those circumstances and purposes, as represented by the agent's maxim. See also Kitcher ("Kant's Argument for the Categorical Imperative," especially pp. 558-60).

expresses this idea in his definition of "practical cognition" as "cognition having to do only with determining grounds of the will" (KpV 5:20). His point, I take it, is not that practical cognition is distinguished from theoretical cognition by its subject matter, but rather that practical reason essentially concerns the determining grounds of the will *conceived as such* and *for the very purpose* of determining the will. Thus, for an agent to represent some maxim as possibly "good" in an act of practical deliberation is to represent it as choice-worthy in an activity the purpose of which is to determine herself to action—to choose.[36]

Often, a maxim will be adopted as the agent's principle because it serves some immediately desired end, as, for example, when I elect to have a slice of cake because I am suddenly craving cake and judge that having a slice of cake is a reasonable pursuit, given my circumstances and wider aims.[37] Of course, a maxim can be adopted on the grounds that it serves some more general aim. For instance, an agent might decide on a particular course of action because she judges that it is conducive to her overall happiness. And an apparently attractive maxim can likewise be rejected because it is judged to be incompatible with broader goals. When my adoption of a maxim forgoes immediate gratification

[36] For a modern defense of this view, see Jeremy David Fix ("Intellectual Isolation").

[37] In order to keep as simple as possible an already complicated chapter, I have remained intentionally vague on the topic of the relationship of empirical desires to maxims. But here I wish to emphasize once again that empirical desires never simply *determine* a transcendentally free agent's action. They can, however, take the form of "incentives" on which she may elect to act. Kantian incentives are rational representations of the putative desirability of pursuing a course of action, which mediate between mere inclinations and what the agent ultimately regards as a reason to act. (Kant defines "inclination" as "the dependence of the faculty of desire upon feelings" (G 4:414n). In Kant's usage, inclinations (*Neigungen*) are the basic sensible desires and impulses.) Consequently, an incentive is a *prima facie* practical reason that guides the agent's adoption of a maxim by specifying an end. For an agent to act on an incentive is to *will* the end or state of affairs specified by that incentive, by taking the necessary and rational means to that end given her circumstances. This interpretation of Kantian incentives follows both Herman ("On the Value of Acting from the Motive of Duty") and Korsgaard ("Morality as Freedom"), of whom the latter asserts that incentives "describe the relation of a free person to the *candidate reasons* among which she chooses" (p. 165, my emphasis). (Korsgaard has since amended her view slightly: she now believes that incentives do not, as such, constitute desires but that desires may form in response to them. Cf. *Self-Constitution*, pp. 122-125.) The claim that an agent is never moved by an incentive except insofar as that agent freely confers on that incentive the normative status of a reason to act is implied throughout Kant's corpus (see, e.g., G 4:412 and KpV 5:79), but is most clearly conveyed in his 1793 work *Religion Within the Boundaries of Mere Reason*. There Kant writes, "freedom of the power of choice…cannot be determined to action through any incentive *except so far as the human being has incorporated it into his maxim* (has made it into a universal rule for himself, according to which he wills to conduct himself); only in this way can an incentive, whatever it may be, coexist with the absolute spontaneity of the power of choice (of freedom)" (Rel 6:24).

for some more overarching pursuit, I demonstrate a capacity to overcome my phenomenologically strongest and most immediate impulses by organizing my practical life according to a diachronically-structured plan. And I engage in this capacity *self-consciously*, with the consciousness, for example, that *I will write chapter three of my dissertation instead of watching Star Wars clips on YouTube.* In doing so, I represent myself as capable of a kind of self-government that is to that extent "free" with respect to my currently monumental desire to watch Star Wars clips on YouTube. Thus, the exercise of this capacity includes, as a constitutive element, a consciousness of myself as exhibiting this capacity for rational self-government.

But as I stated earlier, for an agent capable of pure practical deliberation and action, *all* sensibly-grounded incentives are objects of critical assessment. This is true even of the incentive generated by the agent's calculation of what will serve her happiness over the course of her entire life. In both the *Groundwork* and the second *Critique,* Kant expresses this thought in the claim that the agent's happiness is merely a possible "condition" of action: the agent's conception of her own happiness specifies various ends across a range of circumstances, but the agent pursues those ends in those circumstances only if she wills her own happiness (G 4:416; KrV 5:23ff.). Thus, the exercise of the capacity for prudential thought and action is logically compatible with a type of agency that is incapable of pure practical reasoning. This would be a type of agency that is "conditioned" by sensibility, i.e., one that is incapable of electing to act except on the basis of considerations that are themselves based in sensibility, including those as "remote" from immediate desire as the agent's total happiness.

Nevertheless, the above example goes some way toward explaining, in parallel fashion, how the pure apperception of pure practical deliberation constitutes the agent's consciousness of herself as autonomous. The Formula of Universal Law, as the fundamental principle of pure practical reason, provides a fundamental standard of normative assessment that is *completely independent* of the agent's

sensibility. When an agent adopts a maxim on the basis of that maxim's "lawgiving form," i.e., its suitability for universal adoption as determined by the Formula of Universal Law, she evinces an ability to be motivated by a consideration based entirely in her own pure practical reason. As I argued earlier, the activity of pure practical reason is an essentially self-conscious activity, and the form of this self-consciousness is the form of pure apperception. As such, my engagement in this activity constitutively includes a consciousness of myself as the agent of that very activity—the consciousness of my assessing whether candidate maxims are suitable for universal adoption and determining myself to act at least partly on that basis. Thus, the exercise of this capacity includes, as a constitutive element, a consciousness of myself as exhibiting this capacity for rational self-government and therefore free not just with respect to any one particular empirical incentive but with respect to my sensible nature as such.

As we saw above, Kant claims that we become "immediately conscious" of the moral law "as soon as we draw up maxims of the will for ourselves" (5:29). Kant makes this claim in the context of a more general "Remark" in which he argues that our representation of ourselves as morally obligated is not grounded in an antecedent representation of ourselves as free. For this reason, when Kant claims that we become "immediately" conscious of moral obligation in practical deliberation, he means to highlight the fact that our consciousness of moral obligation is not "mediated" by a prior representation of our freedom. But in addition to this, I believe that Kant's reference to the immediacy of moral consciousness invokes the fact that the moral law is the constitutive principle of autonomous volition. I take his idea to be the following: as autonomous agents, we become conscious of moral requirements "immediately" in practical deliberation because those requirements are criterial of the very activity in which we are engaged. This feature of autonomous agency helps to explain the sense in which empirical practical reasoning is necessarily based on a "condition." The moral law, as the constitutive principle of autonomous agency, provides therein a fundamental normative framework

for deliberation and action. Against this framework, it is a normatively contingent matter whether any candidate maxim serves some empirically-incentivized end. By contrast, whether a maxim has the form of a law is not a normatively contingent matter, but the fundamental criterion for the activity in which we are already engaged.[38]

Let's review what has been claimed so far. I have argued that, on Kant's account, an agent's consciousness of her autonomy is her pure apperception of deliberating and acting according to the Formula of Universal Law, the fundamental principle of pure practical reason. In finite rational agents such deliberation proceeds according to the representation of the imperatival force of morality. But the agent's practical consciousness of the imperatival force of morality is what Kant calls the "fact of reason." Therefore, Kant's "fact of reason" just is the pure apperception of moral deliberation and action, which I claim constitutes the agent's consciousness of her autonomy. This is the sense in which, for Kant, the agent's representation of herself as (positively) free is grounded in the representation of the imperatival force of morality.

Along the way, I have described the relevant form of self-consciousness as the pure apperception of pure practical deliberation *and* action. I have done this because, in the paradigmatic case, pure practical deliberation, i.e., the activity of evaluating candidate maxims for their fitness for universal legislation, culminates in the choice [*Willkür*] of a maxim at least partly on that basis, i.e., the adoption of maxim due in part to its lawgiving form. This reflects the fact that, as I stated above, the activity of practical deliberation is at every stage oriented *toward* action. On this view, action stands to practical deliberation as its essential purpose and natural terminus. But this does not imply that

[38] See also Kant's claim that categorical imperatives "determine the will simply as will" and apply with a kind of necessity that is "independent of conditions that are pathological and only contingently connected with the will" (KpV 5:20). My account suggests an obvious parallel between Kant's assertion of the "immediacy" of moral consciousness in practical deliberation and his claim in the *Groundwork* that whenever the basis for our action is some desired object, "the will never determines itself *immediately*, just by the representation of an action" (G 4:444). Implicit in this is the claim that the Formula of Universal Law is the constitutive standard of action, at least for autonomous agents, and indeed, Kant proceeds to characterize the categorical imperative as "the form of volition as such" (ibid.).

practical deliberation will always culminate in action, much less that pure practical deliberation will always culminate in the choice of a maxim on the basis of its lawgiving form. For this reason, I must clarify and slightly amend my proposal for what the agent's consciousness of her autonomy consists in. In doing so, I will respond to a possible objection based, I allege, in a misreading of Kant's text. This objection makes use of the fact that Kant cites our consciousness of moral requirements as the *ground* of our regarding ourselves as transcendentally free. On that basis, the objection states that we *infer* our transcendental freedom on the basis of our recognition that we are morally obligated, and therefore that the pure apperceptive account of moral consciousness presented in this chapter must be incorrect. In response, I will argue that this objection ignores the larger dialectical aims in light of which Kant makes that claim, and that Kant's account is ultimately subtler than the objection acknowledges.

The first thing to note is that the agent's consciousness of her autonomy does not include, as an essential aspect, any moment of empirical confirmation. Consider the following toy example: first, the agent apperceives the practical-deliberative procedure by which she determines that she is obligated to help others to meet their basic needs;[39] next, she apperceives her choice to determine her will according to the maxim that in such-and-such circumstances she will donate resources to an anti-poverty organization, in order to help alleviate poverty; finally, as she's signing the donation check to Oxfam, she observes her own behavior and thinks, "Here I am, evincing an ability to transcend my sensible nature by motivating myself by the pure thought that some action is morally required!" On the proposal under consideration, each stage in this account—the self-consciousness of pure practical deliberation, the self-consciousness of the will's determination, and finally the "confirmation" of the former stages in the following-through of the intended behavior—jointly constitute the agent's

[39] This is adopted from Kant's fourth example in *Groundwork* II, in regard to the possibility of assessing maxims for universalizability (G 4:423).

consciousness of her autonomy. However, the notion that *any* sort of representation of our autonomy might receive theoretical confirmation through empirical observation is, of course, unambiguously ruled out by Kant's insistence that theoretical cognition is restricted to objects of the natural world, which stand in relations of causal determination. Moreover, Kant states clearly in the Preface to the second *Critique* that practical cognition involves no "extension of [theoretical] cognition to the supersensible" (5:5). So I mention this proposal only to set it aside.[40]

We can nevertheless still ask whether the second "stage" in the above account, the agent's determination of her will on the basis of the lawgiving form of her maxim and the self-consciousness thereof, is a necessary moment in the agent's consciousness of her autonomy. Kant appears to claim, however, that the *mere representation of duty* in practical deliberation is sufficient for regarding ourselves as (positively) free, and this in turn suggests that the agent's consciousness of her autonomy does not include as a necessary component her consciousness specifically of the will's *determination* according to pure practical principles. The relevant text occurs at the end of the "Remark" cited previously. I call attention to this entire Remark, because I will argue that its claims need to be assessed relative to its overriding argumentative aim in order to forestall the objection to my account mentioned above.

Kant provides his Remark relatively early in the first chapter of the *Critique of Practical Reason,* immediately after he argues (under the headings 'Problem I' and 'Problem II', respectively) that, first, only a transcendentally free agent can determine herself to action on the basis of the "universal lawgiving form" of her maxims; and second, that for such an agent *only* the lawgiving form of her maxim could constitute a normatively necessary determining ground of her will (5:29). Since the Formula of Universal Law supplies the standard for the lawgiving form of an agent's maxim, together these claims entail that the Formula of Universal Law sets forth an unconditional practical standard

[40] See also G 4:407, where Kant states that we can never be sure whether we've acted from duty or whether we've acted from a "covert impulse" of self-love.

to all and only those agents who are transcendentally free. Kant's Remark begins with a statement of this thesis: "[transcendental] freedom and unconditional practical law reciprocally imply each other" (ibid.). As we saw, Kant offers a substantively nearly identical claim in the opening paragraphs of the final chapter of the *Groundwork*. There, after providing a short argument, he concludes that "a free will and a will under moral laws are one and the same" (G 4:447).

The purpose of the Remark is to settle a different question: roughly,[41] whether our representation[42] of the imperatival force of morality (and thus our representation of ourselves as morally obligated) is based in a representation of ourselves as transcendentally free, or rather whether our representation of ourselves as transcendentally free is based in a representation of the imperatival force of morality. Of course, we know where the second *Critique* stands on this question: the moral law is the "ratio cognoscendi" of our freedom (5:4n); our representation of ourselves as free has a basis in our practical consciousness of the unconditional authority of the moral law, which Kant calls a "fact of reason." To establish this conclusion, Kant's "Remark" progresses through four main steps:

(i) First, Kant argues that our representation of ourselves as free cannot constitute a basis for regarding ourselves as morally obligated. The argument he provides here is rather brief and unpersuasive. He claims, first, that we lack an "immediate" consciousness of our freedom, since "the first concept of it is negative"; and second, that we cannot "conclude [that we are transcendentally free] from experience, since experience lets us cognize only the law of appearances and hence the mechanism of nature, the direct opposite of freedom" (5:29). The second part of this argument is clear enough: we cannot infer a capacity for transcendental

[41] So rough, in fact, that I think the Remark invites a picture of moral consciousness that was not Kant's own.

[42] I choose the term 'representation' here because Kant refers in this Remark both to (i) our "consciousness" of "a pure practical reason," the latter being "identical with the positive concept of freedom" (5:29) as well as to (ii) our "cognition" of "the unconditionally practical." The term 'representation' is compatible with both of these ways of expressing the question that drives the Remark. Of course, his official view appears to be that our *consciousness* of the imperatival force of morality in practical deliberation—i.e., the "fact of reason"—constitutes our *cognition* of the moral law.

freedom on the basis of the representation of objects that stand in relations of causal determination to each other. However, the first part of this argument raises two questions: (1) Why exactly can't we have an "immediate" consciousness of our negative freedom, that is, one that isn't grounded in our moral consciousness? Kant offers no basis for this claim here, other than to imply that we can only have an immediate consciousness of a capacity in its positive determination. (2) Why can't we have an immediate consciousness of our positive freedom?

I take up the second question in the next section. But with respect to the first question, it bears mentioning that Kant's emphasis on negative freedom in this passage is a dialectical red herring. For, as I argue in chapter four, *even if* we possessed an "immediate consciousness" of ourselves as negatively free (or were otherwise warranted in so regarding ourselves), this would not be a sufficient ground for regarding ourselves as morally obligated, and thus could not serve as the basis of a deduction of the moral law.

(ii) Next, Kant claims that it is "the moral law of which we become immediately conscious" in practical deliberation and proceeds to provide an account of how "consciousness" of the moral law is possible (5:30). I have already discussed the first claim, and I will return to the rest of the account subsequently, as I believe it provides some of the best textual evidence for imputing to Kant the view that moral consciousness is a form of pure apperception.

(iii) At the third step, Kant relates the results of (i) and (ii) to his critical system more generally. Specifically, he claims that it is our capacity for pure practical reason that generates in us the concept of freedom, and that in doing so, it "poses to speculative reason…the most insoluble problem" (ibid.). The problem to which Kant refers is the question of whether there is, in addition to the causality of nature, a causal power that is spontaneous with respect to the laws

of nature. Kant then refers to the fact that this question produces an "antinomy" within speculative reason.[43]

(iv) Finally, Kant concludes this section with a presentation of the "order" of the concepts of freedom and morality in experience, claiming that experience "confirms" that our consciousness of freedom has a basis in our recognition of the moral law.

Step (iv) is where Kant implies that the mere recognition of the authority of the moral law in practical deliberation is sufficient for regarding oneself as positively free. He illustrates this through an example that I quote in full:

> [E]xperience also confirms this order of concepts in us. Suppose someone asserts of his lustful inclination that, when the desired object and the opportunity are present, it is quite irresistible to him; ask him whether, if a gallows were erected in front of the house where he finds this opportunity and he would be hanged on it immediately after gratifying his lust, he would not then control his inclination. One need not conjecture very long what he would reply. But ask him whether, if his prince demanded, on pain of the same immediate execution, that he give false testimony against an honorable man whom the prince would like to destroy under a plausible pretext, he would consider it possible to overcome his love of life, however great it may be. He would perhaps not venture to assert whether he would do it or not, but he must admit without hesitation that it would be possible for him. He judges, therefore, that he can do something because he is aware that he ought to do it and cognizes freedom within him, which, without the moral law, would have remained unknown [*unbekannt*] to him. (5:30)

In this example Kant affirms that as agents who recognize moral requirements we regard ourselves as capable of acting on the basis of our recognition of those requirements, even when they conflict with our inclination toward self-preservation, our "love of life."[44] Moreover, in the context of the wider Remark, this example is intended to lend support to the proposition that our recognition of moral

[43] Kant is referring here to the third antinomy of the *Critique of Pure Reason*, which was briefly discussed in chapter two. See KrV A444/B472ff.

[44] He implies through this example that this representation constitutes cognition, but this should of course not be construed as *theoretical* cognition: he has in mind a distinctively practical form of cognition, such that our entitlement to regard ourselves as free is restricted to the practical domain.

dictates is the *sole* ground of our regarding ourselves as transcendentally free (in either the positive or negative sense).

I have claimed that our consciousness of our autonomy—our consciousness of our *positive* freedom—consists in the pure apperception of pure practical deliberation *and* action. But Kant clearly suggests here that an autonomous agent can cognize his freedom merely on the basis of his recognition of moral requirements, regardless of whether he determines himself to act on that basis: the imagined interlocutor is able to recognize that he possesses the capacity for moral motivation without "ventur[ing] to assert whether he would do it or not" (ibid.).

My interpretation can accommodate this passage by making a small emendation. Strictly speaking, an agent's pure apperception of pure practical deliberation constitutively includes a representation of herself as capable of determining her will according to pure practical principles—and therefore autonomous—*regardless of whether she so determines her will.* Indeed, this is not an *ad hoc* amendment, but is rather implied by the general conception of practical deliberation stated earlier, according to which practical deliberation is at every stage oriented *toward action*, that is, with a view toward the will's determination. To deliberate according to the Formula of Universal Law is, on this conception, not to represent the lawgiving form of some candidate maxim as a good-making feature in a motivationally inert sense—that is, in such a way that does not already bear on the question of what to do—but rather to represent a maxim's lawgiving form as a basis for the determination of one's own will. And to represent any consideration as a determining ground of one's will in the relevant sense is to presuppose a capacity for determining one's will on the basis of that very consideration. Of course, the interlocutor in Kant's example is not actually faced with the circumstance of being asked to give false witness on pain of execution. But he *is* asked to model practical deliberation in thought, to take up the perspective of someone who must act under such circumstances. In doing so, he recognizes a duty not to give false witness, and he recognizes this duty in such a way that it has

direct bearing on his action under those circumstances—i.e., in such a way that represents his categorical obligation as a potential determining ground of his will. Precisely because of this, he also represents himself as capable of determining his will on the basis of his representation of that obligation. The upshot is that the mere representation of duty in practical deliberation constitutively includes the agent's consciousness of her autonomy. Nonetheless, since practical deliberation is always oriented toward action, in the paradigmatic case pure practical deliberation does culminate in the determination of the agent's will. For this reason, going forward I will continue to claim that, on Kant's account, our consciousness of our autonomy is the pure apperception of pure practical deliberation *and* action, with the understanding that 'action' should be understood as the agent's determination of her will, [45] and with the qualification made above.

Although my interpretation can accommodate the example Kant sets forth at the end of the Remark, there is a way of reading this passage that poses a serious challenge to my account, but which I believe is based upon a misunderstanding of the text. For it is possible to read Kant as claiming that we *infer* that we are positively free based upon our representation of moral duty. Indeed, this is a natural way of reading the claim that our imagined interlocutor "judges…that he can do something…because he is aware that he ought to do it" (5:30).[46] An inferential reading of this claim is further suggested by Kant's reference to the "order" of the concepts of freedom and morality, as well as by the implication throughout the Remark that the representation of freedom is "mediated" by a representation of moral obligation. According to this interpretation, the agent's thought process has something like the following form:

[45] That is, 'action' should here be understood as the agent's *determination to act*. In general, we cannot assume that the agent's determination of her will shall be causally efficacious in achieving the end set forth by her maxim. See, for example, Kant's claim that categorical imperatives "determine only the will, whether or not it is sufficient for the effect" (5:20).

[46] The full sentence in German: "Er urtheilt also, dass er etwas kann, darum weil er sich bewusst ist, dass er es soll, und erkennt in sich die Freiheit, die ihm sonst ohne das moralische Gesetz unbekannt geblieben wäre."

(P1) I ought to ϕ.

(P2) I ought to ϕ only if I can ϕ.

(C) Therefore, I can ϕ.[47]

Against this interpretation, I claim that the agent's representation of her autonomy is her pure apperception of pure practical deliberation and action, and thus, that a representation of herself as autonomous is a constitutive feature of her very representation of duty in practical thought. In other words, my interpretation cannot accommodate the proposal that the agent's representation of her autonomy is inferred from her representation of moral duty, because on my interpretation these are not distinct representations, i.e., representations of the sort that could be listed sequentially as in the above depiction of an inferential thought process.

I admit that it is quite natural to read Kant's example on the inferential interpretation. My response is that we should interpret this example in terms of the context of the wider Remark and its overall dialectical aims. The purpose of the Remark is to establish that our consciousness of our freedom has a basis in moral consciousness. Ultimately, this conclusion will serve to justify Kant's claim that a deduction of the moral law cannot be given (5:42ff.). Kant's discussion is therefore intended primarily to underscore the fact that our representation of ourselves as transcendentally free presupposes the representation of moral duty in practical deliberation. To that end, Kant presents the concepts of freedom and morality in terms of their "order" in thought, but we do not need to understand this claim as setting forth the *syllogistic order* in which my cognition proceeds, as if I first represent the moral law as categorically binding, then apply that law to my immediate practical context, and on the basis of my ability to do that, *infer* that I am transcendentally free. In practical deliberation

[47] More precisely, in this context the agent's representation that he "can ϕ" is his representation that he can determine himself to act on the basis of his representation that he ought to ϕ.

I am able to critically assess the various empirical incentives that confront my will, the manifold desires that represent particular courses of action as worthwhile or to-be-pursued. By adopting maxims according to their lawgiving form, I evince an ability to transcend my empirical nature and motivate myself by the thought that a maxim is universally valid. Consciousness of morality is fundamentally the pure apperceptive awareness of the activity whose constitutive principle is the moral law. On my interpretation, the pure apperceptive awareness of this activity *just is* consciousness of my autonomy. Thus, consciousness of positive freedom is identical with moral consciousness.

One might wonder how moral consciousness could be identical to consciousness of positive freedom when Kant explicitly cites cognition of the moral law as the *basis* (the "ratio cognoscendi") of our representation of transcendental freedom more generally. However, it is *extremely* telling that within the Remark Kant introduces this topic with a disclaimer concerning the ensuing discussion: "I do not ask here whether [freedom and unconditional practical law] are in fact different or whether it is not much rather the case that an unconditional law is merely the self-consciousness of a pure practical reason, this being identical with the positive concept of freedom" (5:29). This is terminologically quite awkward, and in his haste to get to the main topic Kant runs together two questions that, at least at this early stage in his exposition, should be distinguished: (1) whether positive freedom is identical to the capacity for pure practical reason, which in finite agents is the capacity for moral agency; and (2) what the *self-consciousness* of our capacity for pure practical reason consists in, and its relation to "unconditional practical law." However, we already know that for Kant positive freedom is our capacity for pure practical reason, and Kant refers to their identity in the above passage.[48] Moreover, the passage virtually demands a rational reconstruction, for it is strictly speaking

[48] The referent of 'this' [*diese*] in 'this being identical with the positive concept of freedom' [*diese aber ganz einerlei mit dem positiven Begriffe der Freiheit sei*] is pure practical reason as mentioned in the phrase 'the self-consciousness of a pure practical reason' [*das Selbstbewusstsein einer reinen praktischen Vernunft*], and not the self-consciousness [*das Selbstbewusstsein*] of pure practical reason. This is evident from the feminine declension of '*diese*'. I am grateful to my former German teacher Emily Jones for confirming this for me.

nonsensical to ask whether "unconditional practical law" (i.e., the fact of our moral obligation) might be identical to a form of self-consciousness. For this reason, I take Kant to be raising the additional question of whether our *representation* of unconditional practical law—that is, our representation of moral duty in practical deliberation—is really "the self-consciousness of a pure practical reason," i.e. the self-consciousness of our "positive freedom." More importantly, I take Kant to be raising this question *in order to set it aside*.

If that is right, then the fact that certain aspects of this Remark invite an inferential reading is not good evidence for that reading. For that means that the entire discussion takes place within the context of Kant setting aside the possibility that our consciousness of positive freedom just is our consciousness of moral deliberation, in order to emphasize the fact that we do not have insight into the nature of our freedom independently of moral consciousness. This helps to explain why Kant implies that our cognition of morality serves as the basis, or ground, on which we regard ourselves as free, when his considered view seems to be a bit subtler: our consciousness of our freedom just is our consciousness of unconditional practical laws, as the pure apperceptive awareness of the activity of pure practical reason guided by those very laws. So in one sense, Kant does assign explanatory primacy to the moral law, inasmuch as it serves as the fundamental principle of the activity the pure apperceptive awareness of which is moral self-consciousness. But in another, he doesn't, since this self-consciousness is constitutive of that very activity.

This interpretation is supported by Kant's own remarks, and is able to clarify and systematize various opaque claims Kant makes about the nature of moral consciousness. One piece of evidence consists in a pair of notes written in the textbook he used for a lecture course on ethics.[49] They were likely written during the years 1776-1778, although they may date to the 1780s.[50] The first of these

[49] *Notes and Fragments*, p. 405.

[50] *Notes and Fragments*, p. 442.

notes principally concerns the relationship between morality and autonomy. There Kant emphasizes that morality requires that "a mere form of actions…have the power of an incentive" and that this indicates the agent's absolute self-determination (NF 19:183). He concludes that "Freedom is apperception of oneself as an intellectual being that is active" (ibid.). In the second he claims that "[t]he apperception of…self-activity is the person" (ibid.).

Some of the best textual evidence for my interpretation is contained within the Remark itself. In particular, I believe that Kant has in mind the pure apperceptive awareness of the activity of pure practical reason, as well as the identity of moral self-consciousness with the consciousness of our positive freedom, in the following passage:

> But how is consciousness of that moral law possible? We can become aware of pure practical laws just as we are aware of pure theoretical principles, by attending to the necessity with which reason prescribes them to us and to the setting aside of all empirical conditions to which reason directs us. The concept of a pure will arises from the first, as consciousness of a pure understanding arises from the latter. (5:30, my emphasis)

This comment is contained within step (ii) of my division of the Remark. It closely follows Kant's claim that we become "immediately conscious" of the moral law when we propose maxims to ourselves (5:29). Kant's identification of moral consciousness with the agent's "attending to…the setting aside of all empirical conditions" resonates with my identification of moral consciousness with the pure apperceptive awareness of pure reason's practical activity. Note the underlined passage in particular, where Kant makes an association between our awareness of pure practical laws and "the concept of a pure will." Given the context of this passage—again, Kant is attempting to show that our consciousness of our freedom presupposes moral consciousness—we should understand by 'the concept of a pure will' the concept of an autonomous (hence transcendentally free) being. The abrupt and unexplained shift in focus from consciousness of moral laws to the genesis of the concept of freedom indicates an implicit identification of consciousness of moral laws with consciousness of freedom. Most importantly of all, I believe this passage contains an allusion to the pure apperception

of our subjectivity in the phrase 'consciousness of a pure understanding'. If we understand by this phrase consciousness of oneself *as subject*, and if the interpretation I advanced in chapter one is correct, then the parallel Kant is noting in the genesis of "the concept of a pure will" and "consciousness of a pure understanding" is exactly the parallel I have been arguing for: just as consciousness of myself *as subject* is the pure apperceptive awareness of rule-governed activities of combination, so consciousness of myself *as an autonomous agent* is the pure apperceptive awareness of the activity of pure practical reason, the rules of which are various categorical imperatives.[51]

Finally, we do see Kant later assert outright what he previously left implicit or vague. First, he defines the fact of reason as "autonomy in the principle of morality by which reason determines the will to deeds" (5:42). He then claims that "this fact is inseparably connected with, and indeed identical with, consciousness of freedom of the will" (ibid., my emphasis). These claims amount to what seems to be Kant's clearest statement that consciousness of freedom is constitutive of the activity of pure practical reason. Finally, in the context of claiming that the concept of freedom "means nothing else" than the concept of pure practical laws in an intelligible world, Kant declares, "How this consciousness of moral laws or, what is the same thing, this consciousness of freedom is possible cannot be further explained" (ibid., my emphasis). My take on the fact of reason provides a systematic way of interpreting these various stray remarks.

§3.3. In the previous section I presented and defended an interpretation of the fact of reason according to which it should be understood as formally analogous to the pure apperception of our mental agency in theoretical cognition. On this interpretation, an agent's consciousness of her

[51] Moreover, by "pure theoretical principles" I take Kant to mean the principles of the understanding. Compare this passage to KpV 5:45, where Kant invokes "principles of that pure speculative reason [that] do no more than make experience possible." "Mak[ing] experience possible" is, of course, precisely the function of the principles of the understanding. Since the mode of pure apperception mentioned in the first *Critique* consists in the subject's awareness of the activity of combining representations *according to principles of the understanding*, Kant's reference to "pure theoretical principles" is further evidence that we should understand Kantian moral consciousness as a form of pure apperception.

autonomy is identical with moral consciousness. This is the sense in which the moral law is the "ratio cognoscendi" (KpV 5:4n) of freedom: my awareness of the latter consists in my pure apperceptive awareness of adopting maxims on the basis of their fitness for universal legislation. I will close this chapter by presenting what I take to be the relationship of this identity to Kant's claim in the second *Critique* that a deduction of the moral law cannot be given.

Before proceeding, it will be helpful to point out that a certain strategy for establishing the moral law is blocked by the fact that the moral law, as given by the Formula of Universal Law, is a *constitutive standard* of autonomous agency. We noted in chapter one that constitutive standards are requirements on being a *kind* of thing, whether that thing is an object or an activity. In particular, the Formula of Universal Law ('FUL') is the constitutive standard of autonomous agency in the sense that it supplies the standard for *what it is* to act autonomously. On this view, autonomy *just is* the activity of adopting maxims on the basis of their fitness for universal legislation, the standard of which is given by FUL. FUL therefore specifies the internal or constitutive principle of autonomous agency, and one cannot engage in the activity of autonomous willing unless FUL is the principle for the determination of one's actions.

If this is right, then it would be evidently circular to try to establish the claim that all rational agents recognize the moral law as an unconditional practical principle, by claiming that all rational agents necessarily uphold and regulate their actions according to an ideal of autonomous agency. For the concept of an autonomous agent is the concept of an agent whose fundamental principle of volition is FUL, so to uphold autonomy as a regulative ideal just is to uphold FUL as the principle of one's volition. FUL is precisely what gives content to our notion of autonomous agency, as its internal standard, so we lack a non-moralized conception of autonomous agency on which basis we could ground our recognition of the moral law. As we will see in the next chapter, it is precisely this feature

that I think Kant alludes to in his famous worry about a "circle" in Groundwork III's initial argument for the reality of moral obligation.

While an argument of this sort is blocked by the fact that FUL is a constitutive standard of autonomy, we might think that our fundamental representation of autonomy is not that of a regulative ideal laden with moral content, but rather the self-awareness that attends autonomous agency. On this view, an agent's representation of autonomy consists, first and foremost, in her consciousness of her own autonomy in practical deliberation and action. And if this is right, we might be able to appeal to such consciousness as the basis of our claim that all rational agents necessarily act under the idea of freedom (understood here as *positive* freedom), and avoid the worry that our making this claim amounts to nothing more than the groundless assertion that every rational agent regulates her actions according to an ideal of moral agency. While the inference from autonomy to moral agency would be quick (it is, after all, an identity), the premise that every rational agent necessarily regards herself as autonomous in practical deliberation and action would be grounded in a phenomenological datum: the autonomy of which each of us is first-personally aware.

Recall from §3.1 that a philosophical "deduction" in general tries to validate our right to apply a synthetic, *a priori* concept or principle by revealing its cognitive origins, and thus the source of its objective validity. We also saw that in the *Critique of Practical Reason* Kant invokes the "fact of reason" in place of any attempt to provide a deduction of the moral law, which he now claims *cannot be given*. By now it should be apparent that the self-consciousness of autonomous agency cannot be made the basis for a deduction of the moral law. For that awareness just consists in the pure apperceptive awareness of the activity of pure practical reason as guided by the moral law, and is thus identical with moral consciousness. The self-awareness of autonomous agency is not a distinct cognitive activity that makes possible our interest in morality. It is rather the case that this self-awareness is the pure apperceptive component of a fundamental activity whose constitutive standard *is the moral law*. Kant's

understanding of moral consciousness thus provides the answer to the question, posed at the end of §3.1, of why our consciousness of freedom is ultimately rooted in our recognition of moral requirements: it is because this consciousness just is the apperception of *moral* deliberation and action.

I have argued that an appeal to autonomy, considered either as a regulative ideal or as something of which we are each first-personally aware, cannot constitute the basis of a deduction of the moral law. We might then wonder whether there is perhaps some other, hitherto unrecognized route to the same conclusion. In fact, I think this possibility is ruled out by the moral law's status as a *fundamental practical principle*, which (I will now show) entails that we could have no insight into its reality or possibility except through the apperceptive awareness that constitutively attends autonomous agency.

As we saw in chapter one, a principle is a rule that guides and determines a rational capacity: the realization of this capacity, considered *qua* rational capacity, just consists in the active deployment of this rule. For this reason, a principle is not merely associated with a capacity, but essential to that capacity's individuation as the kind of capacity that it is. (We can thus think of a principle as the internal or constitutive standard of its respective capacity.) So, for example, the categories of the understanding are each associated with a particular rule of combination, which in turn individuates a particular capacity for representation (for example, the capacity to determine intuitions in regard to subject-predicate relations, to cite the category of substance). Now, the moral law, as given by the Formula of Universal Law, is the fundamental principle of pure practical reason, which entails that there is no more fundamental principle of which the moral law is an instance; as such it individuates and supplies the constitutive standard for a capacity (in this case, the capacity for pure practical thought and action) that cannot be explained in terms of more basic rational capacities. In particular, there is no more basic capacity that stands to pure practical reason as its transcendental precondition, hence no more basic capacity on which basis we could ground its possibility. Moreover, that we

possess this capacity cannot be inferred from or justified on the basis of some other, independent capacity that we possess.[52]

If the capacity for pure practical deliberation and action itself served as a transcendental condition for, or was a constitutive component of, some capacity whose reality was given or otherwise not in doubt, then we could provide a deduction of the moral law by showing that this is the case. (Indeed, this is the form taken by the Transcendental Deduction of the Categories. Kant demonstrates the objective validity of the categories of the understanding precisely by showing that they are conditions of the possibility of self-conscious, empirical representation.) But this is exactly where the moral law's status as a *practical* principle comes to the fore. Pure practical reason is the form that the "faculty of desire," the capacity "to be by means of [one's] representations the cause of the reality of the objects of those representations" (KpV 5:9n), takes in autonomous beings. Therefore, whereas principles of the understanding are conditions on the possibility of the representation of *given* objects, nothing figures in the practical case as the analog of such objects; practical reason is *productive* in regard to its objects, and so the moral law does not and cannot serve as the condition of representations whose object is given. Any particular capacity in which the capacity for pure practical deliberation and action is implicated will produce its object actively in accordance with pure practical principles (for example, in the implicit thought that some action is morally permissible), and we discover the role of pure practical reason not by prescinding from given representations, but through the pure apperceptive consciousness that we are determining our wills according to such principles.

Kant appeals to both the fundamentality of the moral law and its status as practical in explaining why the moral law admits of an "exposition" but not a deduction:

> With the *deduction* [of the moral law]…one cannot hope to get on so well as was the case with the principles of the pure theoretical understanding. … For, the moral law

[52] As I argue in chapter four, this explains the failure of Kant's attempt in *Groundwork* III to justify pure practical reason by appeal to a parallel with theoretical reason's production of ideas. Any structural similarity between the two capacities is entirely orthogonal to the question of the former's possibility.

> is not concerned with cognition of the constitution of objects that may be given to reason from elsewhere but rather with a cognition insofar as it can itself become the ground of the existence of objects and insofar as [pure] reason, by this cognition, has causality in a rational being…
>
> But all human insight is at an end as soon as we have arrived at basic powers or basic faculties; for there is nothing through which their possibility can be conceived. (KpV 5:46-7)

I have already provided what I take to be the correct reconstruction of this passage. Because pure practical reason is a *fundamental* capacity, its possibility cannot be explained by or grounded in an account of some more basic capacity. Because it is a *practical* capacity, there is no given capacity of which it serves as a transcendental condition. And although we are aware that we are autonomous agents, and thus possess a capacity for pure practical thought and action, this awareness is a constitutive aspect of that very capacity, and hence cannot serve as the ground of its deduction. Together, these claims entail that, necessarily, there is no capacity on the basis of which we could deduce that the moral law applies to us. It follows that a deduction of the moral law is impossible.

§3.4. Despite several early attempts, Kant came to believe that this "vainly sought deduction of the moral principle" cannot be given (KpV 5:47). In its place, Kant asserted that our consciousness of moral requirements is a "fact of reason." I have argued that he was right to conclude that such a deduction is impossible. The overarching theme of this chapter is that Kantian moral consciousness must be understood on the model of pure apperception. This model has two distinguishing features. First, it is the consciousness of engaging in a rational activity, i.e., a consciousness that constitutively identifies the apperceiver as the agent of that very activity. Second, it is not a higher-order consciousness, i.e., one that takes that activity as its object, but is rather *internal* to the activity itself. The doctrine of pure apperception found in the first *Critique* was motivated in part by Kant's recognition that the consciousness of one's subjectivity cannot be understood on an empirical-observational model. Since, as I have argued, autonomous agency is an essentially self-conscious

capacity, a formally identical problem arises for its possibility, and the form of Kant's solution is the same. To be conscious of one's autonomy *just is* to be apperceptively aware of one's pure practical deliberation and action. Since the principle of pure practical reason is the Formula of Universal Law, consciousness of one's autonomy is identical to the consciousness of deliberating and acting from moral principles. As a consequence, such consciousness cannot be made the basis of a deduction of the moral law.

However, given the results of chapter two, the pure apperceptive character of moral consciousness suggests a different possibility, namely, that our status as moral agents could serve as an item of reflective self-cognition in a possible critique of practical reason. For as we saw in that chapter, the critique of pure theoretical reason itself proceeds on the basis of items of reflective self-cognition made available to us by the pure apperceptive character of theoretical cognition. In other words, if the account of critique offered in chapter two is correct, Kant's reliance on a distinctively *practical* form of reflection in a critique of practical reason would not by itself constitute any reversal of method. To the contrary, such reliance would be an *expression* of the methodological commitments of critique, inasmuch as the presumption of a capacity for reflective self-cognition is grounded in the presumption of our rational autonomy. I defend such an interpretation in the next chapter.

CHAPTER FOUR

Reflection and *Critique* in Kant's Practical Philosophy: Why a Deduction of the Moral Law Is Not Necessary

§4.0 The previous chapter argued for two main claims: (1) that an agent's consciousness of her autonomy consists in her pure apperception of moral deliberation and action; and (2) that because of this, such self-consciousness cannot be made the basis for a deduction of the moral law. According to this view, moral consciousness—the representation of the imperatival force of morality in practical deliberation—constitutively includes the agent's consciousness of herself as an autonomous, or *positively* free, being—i.e., her consciousness of herself as able to determine her will on the basis of pure reason alone. The foregoing characterizations emphasize different aspects of what is a single capacity: the capacity to choose maxims on the basis of their lawgiving form, the standard of which is the Formula of Universal Law.

In the *Critique of Practical Reason* (1788), Kant appeals to such consciousness, which he labels a "fact of reason," in place of any attempt to provide a deduction of morality such as he earlier attempted in the third chapter of the *Groundwork of the Metaphysics of Morals* (1785). Accordingly, Kant's later position is not just that a deduction of the moral law *cannot* be given but that it *need not*: even in the absence of such an argument, we rational agents are fully justified in regarding ourselves as beings who face moral obligations, and we are justified because we are conscious of the imperatival force of morality in first-order practical deliberation. In what follows I will provide an interpretation and partial defense of this claim. I say "partial" because my conclusion will be that Kant's later position does not constitute a regression into dogmatism, and therefore is at least consistent with the standards of Kantian critique. Indeed, Kant's disavowal of the need for a deduction of the moral law is not only fully consistent with, but in fact rooted in, the core methodological commitments of critique. Consequently, Kant's reliance on a "fact of reason" in place of a deduction does not constitute a

departure from the method of critique, much less a slide into pre-critical dogmatism, as even sympathetic commentators have sometimes worried.

The plan for this chapter is as follows. In §4.1, I provide a more detailed expression of the worry that Kant's invocation of a "fact of reason" constitutes a departure from his critical method. In §4.2, I argue that the doctrine of the fact of reason indicates Kant's reliance on a distinctive form of reflection, which I label 'practical reflection', and that such reliance marks a point of continuity with the method of the first *Critique*. This argument refers back to chapter two, where I provided an account of the methodology of critique according to which critique essentially involves reflection on one's own cognitive activity. Once this aspect of critique is recognized, it will become clearer why Kant can think, without lapsing into dogmatism, that pure practical reason "proves its reality" by its activity and that no special critique of *pure* practical reason is required. The latter issue is addressed more fully in §4.3, where I connect this claim to the idea that critique is an expression of reason's normative autonomy. In §4.4, I provide an account of the deduction that Kant attempts in *Groundwork* III, arguing that while this argument likewise depends on practical reflection, it also departs from the core methodological commitments of critique in important respects.

§4.1 Even among sympathetic interpreters, Kant's reliance on a "fact" that grounds our self-conception as moral agents is sometimes thought to reflect his having regressed into precisely the kind of pre-critical dogmatism that he criticized in the first *Critique*.[1] Among early critics, Schopenhauer was especially hostile to Kant's later position: "practical reason with its categorical imperative appears

[1] John Rawls remarks that the doctrine of the fact of reason "has looked to some like a step backward to some kind of intuitionism or else to dogmatism" (*Lectures on the History of Moral Philosophy*, p. 268).

more and more as a hyperphysical fact, as a delphic temple in the human soul. From its dark sanctuary oracular sentences infallibly proclaim, alas! not what *will*, but what *ought* to happen."[2]

It certainly *seems* uncritical, and therefore deeply un-Kantian, to account for the truth of a synthetic, *a priori* proposition by stipulating a "fact"—if indeed that's what Kant is doing. As an argumentative maneuver, it brings to mind Bertrand Russell's famous quip that postulating what we want has all the advantages of "theft over honest toil."[3] Even if we grant that a deduction of the moral law is impossible, it seems that we should resist on Kant's own methodological grounds any suggestion that the moral law—or, more precisely, the proposition that the moral law applies to the human will as its fundamental principle—is for that reason exempt from critical examination and can therefore serve as a foundational premise in practical philosophy. One might reply that this objection overlooks Kant's restriction of theoretical knowledge to the empirical domain, and that the doctrine of the fact of reason is licensed by the fact that theoretical reason could not be justified in denying our freedom, since it lacks all insight into the nature of things-in-themselves. But this wrinkle in the account fails to address the central worry, since lacking grounds for morality's denial does not entail possessing grounds for its affirmation. In order to give shape to this worry, I will now attempt to state more precisely what is so apparently dogmatic about Kant's later position—and correct one natural misinterpretation.

To begin, what exactly is meant by calling a position "dogmatic" or "uncritical"? In the B-edition Preface to the first *Critique*, Kant characterizes "dogmatism" as "the presumption of getting on solely with pure cognition from (philosophical) concepts according to principles, which reason has been using for a long time without first inquiring in what way and by what right it has obtained them. Dogmatism is therefore the dogmatic procedure of pure reason, **without an antecedent critique of**

[2] Qtd. in Allison (*Kant's Theory of Freedom*, p. 230).

[3] *Introduction to Mathematical Philosophy*, p. 71.

its own capacity" (KrV Bxxxv). In this context, Kant understands the "dogmatic procedure" to be any theoretical investigation that relies on synthetic, *a priori* principles without examining the legitimacy or applicability of those principles. Such an investigation accepts the truth or legitimacy of certain "pure" (*a priori*) principles without first inquiring into reason's own capacity to apply those principles in such a way as to yield knowledge. For example, among the procedures Kant rejected were the various rationalist attempts to infer the existence of an Ultimate Cause on the basis of the Principle of Sufficient Reason. These attempts were "dogmatic" in that they proceeded without first establishing their "right" to deploy this principle and, more generally, without first engaging in critique, which in this context Kant emphasized as a "preparatory activity necessary for the advancement of metaphysics as a well-grounded science" (KrV Bxxxvi). Kant's own critical investigations would yield the conclusion that the Principle of Sufficient Reason is a "ground of possible experience" and thus valid insofar as it applies to objects of empirical knowledge, but that its scope does not extend beyond such objects; in particular, it does not furnish us with knowledge of the ultimate nature of things (KrV A201/B246). So the dogmatist's "presumption of getting on solely with pure cognition" is objectionable not just because it proceeds without first establishing the legitimacy of its principles, but moreover because it leads to genuine error, by supposing that principles that properly apply only to empirical objects can yield insight into things-in-themselves. That we do not have knowledge of things-in-themselves is, of course, one of the major tenets of Kant's critical philosophy.

This last point suggests one way in which the doctrine of the fact of reason might be thought to be dogmatic. For the Kantian restriction of the objects of our knowledge to empirical objects would plainly be violated if Kant is now claiming that we possess an extrasensory perception of an independent moral reality. Perhaps this is what Schopenhauer was getting at with his admonition that "practical reason with its categorical imperative appears more and more as a *hyperphysical* fact."[4] The

[4] Qtd. in Allison (*Kant's Theory of Freedom*, p. 230), my emphasis.

suggestion might be that, although Kant carefully demarcates the boundaries of knowledge for theoretical purposes, his own practical philosophy presupposes a capacity for theoretical knowledge that transcends those very boundaries. Moreover, by stipulating a "fact" and denying that a critique of pure practical reason is even necessary, he makes no attempt to justify or explain what exempts practical reason from this restriction or what accounts for such cognitive capacities.

On this interpretation, Kantian moral knowledge is to be understood on a quasi-perceptual model. And since Kant claims in the second *Critique* that our recognition of the moral law is the basis of our regarding ourselves as free, this suggests that the fact of our transcendental freedom is something we *infer* only after we "perceive" that we are morally obligated (and hence the kind of beings to whom moral standards apply). I argued in the previous chapter that this is precisely *not* how a moral agent recognizes the authority of the moral law or cognizes her own freedom—again, this is entailed by moral consciousness being a kind of pure apperception—but it suffices to say that interpretive charity and Kant's own words tell against such a reading. It is implausible that Kant, who in the first *Critique* was so adamant that we lack theoretical knowledge of noumenal reality, would make such knowledge the foundational element of his practical philosophy. Furthermore, although Kant in his discussion of the fact of reason claims that the moral law "forces itself upon us" (KpV 5:31), which might be taken to mean that moral facts are "given" to us in a quasi-perceptual manner, he immediately clarifies this remark. *If* it were assumed that we are autonomous, it would analytically follow that we are morally obligated. But Kant explains that to arrive at the fact of our moral obligation in this way, an "intellectual intuition would be required," which, according to Kant, "certainly cannot be assumed here" (5:31). 'Intellectual intuition' is Kant's label for a kind of "immediate" perception—that is, conceptually unmediated representation—of noumenal reality. In the first *Critique* Kant claims that we would need to have a power of intellectual intuition in order to have knowledge of noumena; however, Kant introduces the idea of this mode of cognition specifically to distinguish it from our

own (KrV B307). It is because we are discursive subjects, subjects who rely on both concepts and sensible intuitions for our theoretical cognition, that we are incapable of cognizing things-in-themselves. In light of these commitments, we can see that Kant rejects the assumption that we have an intellectual intuition of our freedom because he more generally rejects the idea that we possess *any* capacity for intellectual intuition. (Admittedly, one wishes Kant had been clearer on this issue, since all he says in connection to our positive freedom is that an intellectual intuition "cannot be assumed.") Since such intuitions were a prerequisite for knowledge of noumenal reality, Kant is not now claiming that the fact of reason constitutes a quasi-perceptual awareness of a "hyperphysical" (i.e., noumenal) state of affairs. If the doctrine of the fact of reason is dogmatic, it is not because Kant's practical philosophy transgresses the boundaries established by his theoretical philosophy.[5]

Nevertheless, Kant does claim that we are conscious of a "fundamental law" (namely, the moral law presented as the categorical imperative) and that its applicability to us *does not need to be justified* (5:47). With the latter claim Kant renounces not only the need to provide a "deduction" of the moral law but more generally any attempt to provide a "critique" of pure practical reason (5:3; 5:42ff.). Kant ultimately defends the reality of objective morality and transcendental freedom on the basis of an unargued-for "fact," namely, our consciousness of the unconditional authority of the moral law. However, against Kant's insistence that the moral law "has no need of justifying grounds" (5:47), what reason do we have to think that the fact of reason doctrine amounts to anything more than the groundless assertion of a synthetic, *a priori* principle? On such an interpretation Kant's practical philosophy would amount to the kind of uncritical enterprise for which Kant criticized the dogmatic metaphysicians who preceded him. For suppose Kant had never woken from his famous dogmatic

[5] Of course, part of the explanation for why Kant's practical philosophy does not transgress his theoretical philosophy consists in the fact that Kant makes a formal distinction between practical and theoretical reason, and insists that the major conclusions of his practical philosophy are restricted to reason in its practical use. Cf. G 4:448 and KpV 4:4-5. In §4.3 I provide an account of what this restriction amounts to.

slumber, and had insisted against Hume's skeptical attack that it is a "fact of reason" that every event has a cause. This would *clearly* count as a "dogmatic" response by Kant's own lights. Yet without further explanation, we are left to wonder why such a response is illegitimate in the case of the principles of the understanding but licensed in the case of the fundamental principle of pure practical reason.

Among contemporary commentators, Karl Ameriks has perhaps been the most emphatic in claiming that Kant owes us a positive argument for the reality of moral obligation. According to Ameriks, Kant's abandonment of such a project and concomitant adoption of the doctrine of the fact of reason constitutes a slide into pre-critical dogmatism. In "Kant's Deduction of Freedom and Morality," Ameriks objects to the later doctrine on the grounds that nothing short of a "theoretical" argument for the will's autonomy could provide the philosophical justification of morality Ameriks takes Kant to need. By Ameriks's lights, the *Groundwork*, unlike the second *Critique*, at least acknowledges and tries to meet this demand. Central to this line of thought is Ameriks's contention that the third chapter of the *Groundwork* contains a purely theoretical argument for its conclusion: in contrast to the second *Critique*, the *Groundwork* attempts to provide a "strict deduction" of the objective validity of the moral law, here defined as "a 'linear' argument intended to be logically sound with premises that are all and only theoretical as opposed to practical in any Kantian moral sense."[6] Whatever the failings of the *Groundwork* argument, it "at least sees what Kant's philosophy needs with respect to freedom and…tries to meet that need. It does not rest content with the mere logical possibility of transcendental freedom nor…with saying that we can act as if we are free or as if there is some basis, but not a strong theoretical one, for saying that we are free."[7] Worse still, Kant makes no attempt in the second *Critique* to justify the claim that a deduction of morality need not be given,

[6] "Kant's Deduction of Freedom and Morality," p. 53 fn. 2.

[7] Ibid., p. 65.

and his adoption of the doctrine of the fact of reason in place of any such attempt is a deliberate shirking of the critical demand:

> Kant…totally abandoned the distinctive points of the last section of the [*Groundwork*]. However, in this abandonment Kant gave up not only some vulnerable points but also some valid ones. That is, Kant never gave any reason for ignoring his earlier idea that freedom should be argued for theoretically and that a mere absence of grounds against it is not adequate given the "perplexity" our transcendental freedom presents for the system of speculative knowledge. Instead at the very end of his career Kant frankly acknowledged that his practical philosophy was "dogmatic" and that only his theoretical philosophy was to be called Critical.[8]

Finally, while Ameriks is careful to note that, on Kant's account, our consciousness of the moral law does not take the form of an intuition of a given object, this is only because of the "technical peculiarit[y]" of the fact that the moral law is a principle rather than a particular. Nevertheless, inasmuch as Kant's practical philosophy is grounded in the "nonnaturalistic ultimacy" of the fact of reason, it is in all relevant respects epistemologically on a par with rational intuitionism. For this reason Ameriks takes Kant to be partly to blame for the return ("at least in Germany") of dogmatic metaphysics and the "mystical excesses" of the idealists who succeeded him.[9]

Ameriks represents a particularly clear and forceful distillation of the view that I will be arguing against, so I will be returning to him throughout this chapter. For my purposes, the main items of his attack are the following:

(1) Kant's invocation of a fact of reason in place of a deduction of the moral law is inconsistent with the method of critique inasmuch as it constitutes a form of dogmatism.

(2) Kant owes the reader a theoretical argument for the reality of moral obligation.

(3) The argument of *Groundwork* III acknowledges the demand to provide a theoretical argument, and attempts to provide one.

---

[8] Ibid., p. 72.

[9] Ibid.

In what follows, I will be responding to each of the claims. In later sections, I will show that (2) and (3) reflect a conception of the critical enterprise starkly at odds with the one I articulated at the end of chapter two, where I claimed that critique itself presupposes and exhibits a kind of rational autonomy. As I will argue, the existence of pure practical reason does not require any kind of *theoretical* validation, and the *Groundwork* was not attempting to provide one. Indeed, the very demand for theoretical validation is incompatible with the idea that pure reason itself supplies the standard for its own critique. Thus, on my reading, (2) and (3) reflect a sort of "heteronomous" conception of the sort of justification that Kant needs, by importing a standard of justification antithetical to critique as an autonomous enterprise.

Before taking up these issues, however, I will first argue against (1), the claim that Kant's invocation of a "fact of reason" represents a departure from critical methods. According to my interpretation, Kant's reference to the "fact" of our moral consciousness demonstrates his reliance on a distinctively *practical* form of reflection. Since, as I argued in chapter two, the *Critique of Pure Reason* likewise appeals to items of self-cognition acquired only through first-person reflection on our own cognitive capacities, Kant's continued reliance on reflection in the second *Critique* is at least consistent with the earlier method.

§4.2 The *Critique of Practical Reason* begins with an announcement and something of a disclaimer regarding the aims and ambitions of the book:

> Why this *Critique* is not entitled a *Critique of Pure Practical Reason* but simply a *Critique of Practical Reason* generally, although its parallelism with the speculative seems to require the first, is sufficiently explained in this treatise. It has merely to show *that there is pure practical reason*, and for this it criticizes reason's entire *practical faculty*. If it succeeds in this it has no need to criticize the *pure faculty itself* in order to see whether reason is merely making a claim in which it presumptuously *oversteps* itself (as does happen with speculative reason). For, if as pure reason it is really practical, it proves its reality and that of its concepts by what it does [*durch die Tat*], and all subtle reasoning against the possibility of its being practical is futile. (KpV 5:3)

We see that one of the tasks of a critique of practical reason is to show that there is *pure* practical reason, in the sense of showing that the human will is capable of determining itself to action on the basis of pure rational principles. As I explained in the previous chapter, the implied contrast is with a faculty of practical reason that is sensibly or "empirically" conditioned, that is, one that must draw on non-rational sources of motivation such as are provided by our various sensible inclinations and drives.[10]

In the text quoted above, Kant makes a series of related assertions that will be the focus of this and subsequent sections:

> First, Kant claims that if it can be shown that there is pure practical reason, then it will not be necessary to submit pure practical reason to a special critique.
>
> Second, Kant claims that the object of such a critique would be to determine whether reason's claim to possess a "pure practical" capability is really just an instance where reason "presumptuously *oversteps* itself," as occurs, for example, in exercises of theoretical reason that seek to describe noumenal reality.
>
> Third, and finally, Kant claims that we do not have to worry about reason "overstepping itself" in this case precisely because a capacity for pure practical reason would be a *practical* capacity: as such, it would "[prove] its reality and that of its concepts by what it does [*durch die Tat*]" (KpV 5:3).

Kant's assertion that pure practical reason does not require a critique is a major reversal of the position he espoused three years earlier in the *Groundwork*, which, as we saw in the last chapter, did hold such a critique to be necessary. In the Preface to the *Groundwork*, Kant identified the task of providing a

[10] See for example KpV 5:15.

"foundation of a metaphysics of morals" (the express aim of the entire work) with that of a critique of pure practical reason (G 4:391).[11] It is also evident that at the time of the writing of the *Groundwork* Kant had no intention of producing a separate critique devoted exclusively to practical reason, as he believed that the *Groundwork* sufficed for what the aim of such a critique would be.[12] The third section in particular, titled "Transition from metaphysics of morals to the critique of pure practical reason," was to make good on the advertised promise of securing a philosophical justification of morality. Thus, some such about-face in regard to the kind of argument that could be on offer seems undeniable.

The full significance of the quoted passage will become apparent as we proceed. For the time being, we should note that it gestures toward *some* explanation of Kant's reversal, and thus should serve as a clue as to why Kant believed that a deduction of the moral law is not necessary. Ameriks, we saw, claims that "Kant never gave any reason for ignoring his earlier idea that freedom should be argued for theoretically."[13] Even if we grant Ameriks's assumption that Kant's previously attempted deduction was intended as a theoretical argument, the very opening lines of the second *Critique* suggest the broad contours of an explanation of Kant's later position: namely, that pure practical reason *is practical*, and as such, establishes itself *durch die Tat*.

Less obviously, Kant's claim that pure reason proves its practicality *durch die Tat* is an allusion to the fact of reason. The most literal translation of '*durch die Tat*' is 'through the deed' or 'by means of the deed'. Thus, what Kant is claiming in this passage is that pure practical reason proves its reality through the deed or activity of pure reason. But 'deed' is an alternative translation of '*Faktum*' (indeed,

[11] Kant qualifies this remark a few sentences later when he asserts that a complete critique of pure practical reason would "present the unity of practical with speculative reason in a common principle," which is not a goal of the *Groundwork* (G 4:391).

[12] G 4:391.

[13] "Kant's Deduction of Freedom and Morality," p. 72.

the primary meaning of '*Faktum*' in eighteenth-century German[14]), and Kant himself uses the Latin cognate '*factum*' to refer to the notion of a deed (*Tat*).[15] This implies that 'deed of reason' is an alternative (and quite possibly superior) translation of '*Faktum der Vernunft*', and for this reason many recent commentators have argued that Kant's fact of reason is best understood as a kind of "deed" or activity.[16]

Given what I say here, it might seem that I am likewise committed to such an interpretation. However, it is ultimately incidental to my exegetical aims whether we say that the fact of reason is (i) the *fact* of my consciousness of the unconditional authority of the moral law or (ii) the *activity* of pure practical reason. For, on the interpretation I advanced in the previous chapter, consciousness of myself as the agent of the activity of pure practical reason is constitutive of that very activity, and (for sensibly-affected agents such as ourselves) to be conscious of this activity is to be conscious of oneself as deliberating within moral constraints, and thus standing under moral laws. If this interpretation is correct, then if it should turn out that Kant's usage of '*Faktum*' is better translated as 'fact', the supposed "fact" of moral consciousness turns out to be identical to a form of self-consciousness that constitutively figures in pure practical deliberation and action—and, crucially, it is only by engaging in the activity of pure practical deliberation and action that we first regard ourselves as positively free agents who stand under moral laws. Going forward I will assume that the *fact* of moral consciousness is a finite agent's pure apperception of her pure practical *activity*.[17] The important point is this: since

---

[14] Kleingeld ("Moral Consciousness and the 'Fact of Reason'," p. 62).

[15] MM 6:224.

[16] See, for example, Marcus Willaschek ("Die Tat Der Vernunft: Zur Bedeutung der Kantischen These vom 'Faktum der Vernunft'"); Paul Franks (*All or Nothing: Systematicity, Transcendental Arguments, and Skepticism in German Idealism*, pp. 277-279); David Sussman ("From Deduction to Deed: Kant's Grounding of the Moral Law," pp. 67-68); and Stephen Engstrom ("Introduction," *Critique of Practical Reason* (Pluhar translation), pp. xli-xliii). Against this emerging consensus Pauline Kleingeld has argued that the *Faktum der Vernunft* is best understood as the fact of moral consciousness, but that such moral consciousness is generated out of the activity of reason ("Moral Consciousness and the 'Fact of Reason'"). My own view is closest to Kleingeld's.

[17] See chapter three, footnote 9.

Kant holds that the reality of pure practical reason is disclosed *durch die Tat*, Kant's use of the phrase '*durch die Tat*' plainly anticipates the doctrine of the fact of reason.

This connection is underscored by the fact that Kant's disavowal of the need for a critique of pure practical reason is a concomitant rejection of the need for a deduction of the moral law. (In the *Groundwork,* the task of a "deduction" of the moral law was subsumed under the task of a "critique of pure practical reason" (G 4:446ff.).) Recall that critique serves, positively, to establish the legitimacy of some rational principle or claim to knowledge and, negatively, to set boundaries on reason's legitimate use. While both aspects of critique are represented in the passage quoted above, Kant refers to critique's positive function when he says that pure practical reason does not require a critique because its reality is established *durch die Tat* (5:3). But since critique fulfills its positive function through the method of *deduction*, we can read Kant as claiming that the "deed" of pure reason obviates the need for a deduction. And since Kant invokes the doctrine of the fact of reason in place of any attempt to provide a deduction of the moral law (KpV 5:47-48), this suggests that Kant's claim that pure practical reason has no need for a critique because its reality is revealed through a "deed" anticipates and is ultimately inseparable from his claim that a "fact of reason" supplants the need for a deduction of the moral law. The interpretation I advanced in the previous chapter provides a way of spelling out this connection: the "deed" that Kant has in mind is the activity of pure reason in its practical deployment, which constitutively includes the agent's consciousness of herself as engaged in that very activity. Since the principle of this activity is the moral law itself, as given by the Formula of Universal Law, this agential self-awareness is attended by recognition of the authority of moral principles, and is thus a form of moral consciousness.

Put less abstractly and in the first person: when I deliberate about what to do, I do so according to certain recognized moral principles and within certain recognized moral constraints. But since this is a self-conscious activity, a certain *self*-recognition, the recognition that *I* stand under such-and-such

obligations, is internal to that very activity. Moreover, it is self-conscious in the sense that I recognize *myself* as the agent of the activity in question, and thus as capable of determining myself to action on the basis of my recognition of moral obligations. But these aren't two different forms of self-recognition; rather, internal to moral deliberation is the unified awareness of myself as an agent who stands under—and is capable of acting from—moral principles. And this self-awareness is not made available through introspection but is rather the apperceptive awareness that constitutively figures in my deliberation concerning what to do. Seen in this light, Kant's disavowal of the need for a deduction of the moral law (or a critique of pure practical reason) amounts to the claim that the availability of this form of self-consciousness entitles me—in a restricted sense that will be explained—to regard myself as a transcendentally free moral agent. In other words, internal to the apperception of my practical deliberation is a consciousness of myself as an autonomous agent who faces moral obligations, and precisely because I possess this consciousness I am under no requirement to provide additional justification for these claims through a special critique of my pure practical faculty. (Or so Kant claims.)

I present this account in the first person in order to emphasize the role of *reflection* within the *Critique of Practical Reason*. Recall from chapter two that 'reflection' (*Überlegung*) had several (possibly connected[18]) meanings within Kant's critical system. In one of Kant's usages, 'reflection' denotes an activity of self-cognition made possible by pure apperception, and which gives us, the pure apperceiving subjects, a foothold into the formal logical and transcendental structure of our theoretical cognition. In this way, reflection provides us with the substantive self-cognition that makes a critique of pure theoretical reason possible. So, for example, by reflecting on our discursive activities and their role within our cognitive economy, we can make explicit for ourselves the fundamentality of *synthesizing* representations into a unified whole with respect to the possibility of *analyzing* that unified

[18] See chapter two, footnote 17.

representation in terms of its constituent elements. Relatedly, reflection can disclose to us the formal structure of pure apperception, helping us to identify the role of pure apperception in, for example, categorially-structured empirical judgment.

By linking Kant's doctrine of the fact of reason to his claim that pure practical reason proves its reality through its own activity (*durch die Tat*), we highlight the fact that within the argument structure of the second *Critique*, the "fact of reason" occupies the role of our most fundamental cognition of our practical capacities, as well as the basis for systematic philosophical inquiry. But as I have argued, this cognition is a kind of pure apperception. This entails that the *Critique of Practical Reason* itself proceeds on the basis of considerations we become aware of only through the pure apperception of pure practical deliberation and action. And since reflection was defined in chapter two as a form of pure apperceptive cognition of our various cognitive capacities—apperception being internal to the operation of those very capacities—that is just to say that the second *Critique* avails itself of items of self-cognition made available through reflection on our (pure) *practical* capacities. We summarize this in the thought that the *Critique of Practical Reason* is made possible through *practical reflection*.

Given the theory of reflection advanced in chapter two, this points to a more general truth about the critical method: necessarily, reason's critical self-examination, whether of theoretical or practical reason, draws on reflection as a source and basis for further theorizing.[19] This has two immediate implications that count against Ameriks's charge of dogmatism. The first is that, contra Ameriks, Kant does not take the "mere logical possibility of"[20] or "mere absence of grounds against"[21] transcendental freedom as a sufficient basis for justifying the claim that we are transcendentally free.

[19] Although Henrich focuses on the first *Critique,* he likewise offers his interpretation as a more general account of the critical method. For example, he claims that the general relationship of reflection to deduction is maintained in the second *Critique's* deduction of freedom ("Kant's Notion of a Deduction and the Methodological Background of the First *Critique*," p. 43).

[20] Ameriks ("Kant's Deduction of Freedom and Morality," p. 65).

[21] Ibid., p. 72.

Rather, Kant holds that moral consciousness itself supplies a (non-inferential) basis for regarding ourselves as capable of acting from the representation of moral duty, and thus as positively free. This brings us to the second implication. The doctrine of the fact of reason is epistemically grounded in practical reflection, just as the critique of theoretical reason itself relies on facts and principles made available through reflection on the capacities operative in theoretical cognition. Thus, Kant's reliance on practical reflection is fully consistent with a suitably generalized conception of Kant's critical method, and therefore does not by itself constitute any reason to think that the doctrine of the fact of reason represents a lapse into dogmatism.

A further comparison with the method of the *Critique of Pure Reason* will serve to highlight this last point. I have been arguing that the fact of reason should be understood as a form of pure apperception, formally analogous to the pure apperceptive consciousness that necessarily attends my representations insofar as those representations are "something for me," their subject, and can figure as contents in judgments of the form *I think x*. Call this the *formal* analogy between the fact of reason and the doctrine of pure apperception found in the first *Critique*. I am now claiming a further, but related, analogy between them: the pure apperception of pure practical reason, and the pure apperception articulated in the Transcendental Deduction of the first *Critique*, function in broadly *methodologically* similar—though, as we shall see, not entirely parallel—ways within their respective critiques. For both make possible reflective modes of cognition that ground and orient their respective critiques by generating principles on the basis of which philosophical inquiry can be undertaken. In the Transcendental Deduction, for instance, Kant argues that pure apperception is a necessary condition of our representation of the unity of an object. But the proposition *that* we can represent the unity of an object and *that* such representational unity is made possible by a capacity for pure apperception, are items of self-cognition grounded in first-person reflection on those very capacities. Similarly, the doctrine of the fact of reason is grounded in reflection on our pure practical capacities,

and is invoked by Kant as a basis for regarding ourselves as transcendentally free (5:47-48). To be sure, there are striking differences between these doctrines and their significance within Kant's system. (For example, only the former is invoked to provide a metaphysics of the natural world.) Nevertheless, both doctrines are grounded in reflection on the relevant capacities. That is, in both cases, the relevant arguments get a purchase on us, the readers, through our own capacity for pure apperception, whether of the general logical form of our thought, the transcendental conditions of our representation of objects, or the imperatival structure of our practical deliberation. Consequently, the fact that the second *Critique* proceeds from practical apperceptive consciousness is by itself no more dogmatic than the fact that the transcendental deductions of the categories specifies a role for pure apperception in the operations of the understanding.

§4.3 The aim of the previous section was to show that Kant's invocation of a fact of reason indicates his reliance on practical reflection, which, as such, does not constitute a lapse into uncritical dogmatism. The reader may be wondering whether I am overstating the methodological parallels. In §4.1, I posed a challenge to the doctrine of the fact of reason: any satisfactory defense of this doctrine will have to explain why Kant was not entitled to declare it a "fact of reason" that every event has a cause. The worry is that Kant's invocation of moral consciousness amounts to his asserting that things *are* a certain way merely because they *seem to us* to be a certain way, much as it might initially seem to us (prior to critique, that is) that every event has a cause. Kant claims that pure practical reason, *as practical*, "proves its reality" through its own activity (KpV 5:3). But we have yet to specify what is distinctive about pure practical reason in a way that explains why it does not require a special critique.

This worry is related to, and informs, the notion that we would need a *theoretical* argument for moral obligation in order to be justified in claiming that we are morally obligated. For it is natural to suppose that the purpose of theoretical reason is to describe, and justify to us, the way that things (in

the broadest possible sense of that term) *are*. And this attitude in turn lends itself to a particular conception of critique, according to which critique is an exercise of *theoretical* reason, applied to the special case where the "objects" of theoretical inquiry are our own powers of theoretical and practical cognition. This attitude is reflected in, for example, Ameriks's insistence that Kant owes us a theoretical argument for moral obligation, and that in the *Groundwork*, he attempts to give one.

My plan for this section is as follows. First, I will expand upon the significance of Kant's distinction between theoretical and practical reason. On the basis of the account I provide, I will explain how Kant's defense of the reality of moral obligation is *not intended to have theoretical significance*. Theoretical claims are claims about objects that are external to our own powers of conceptual representation, i.e., our power of judgment. By contrast, our entitlement to regard ourselves as autonomous is an entitlement to regard ourselves as possessing a particular capacity for practical judgment; that we are so entitled entails nothing whatsoever about objects that are external to our representational powers. As I shall argue, it is precisely for this reason that the categories of the understanding require a deduction while "the moral law," i.e., the claim that we are morally obligated, does not.

Along the way, I will provide a diagnosis of why it might seem that a "theoretical" argument for moral obligation is required. In connection with this, I will argue that the demand for theoretical justification is at odds with the normative autonomy of Kantian critique. This demand mistakes our claim to regard ourselves as morally obligated for a description of an external, objective reality, and posits that we are normatively answerable to that reality in making that claim. But this misconstrues the nature and source of our entitlement. In place of this "realist" conception of justification, it's rather the case that the source of our entitlement is the normative autonomy of reason. In particular, the presupposition of our rational autonomy in turn presupposes that we are conscious of the nature of our own rational activity—that we are conscious at least of what *we* are subjectively doing (in acts

of combination, determinations of the will, and so forth), notwithstanding the possibility that the nature of the object or manifold on which we are acting (or through which our representational powers are drawn into operation) eludes us. Thus, we are entitled to assume a power of reflective self-cognition—exhibited, for example, in our claiming to possess a capacity for pure practical deliberation and action—and this entitlement is an expression of our rational autonomy. (Here I make use of §2.3, where I argued that the presumption of a capacity for reflective self-cognition is internal to critique's necessary presumption of the autonomy of reason.)

Let's begin with the distinction between theoretical and practical reason. Kant marks such a distinction early in the Introduction to the second *Critique*, which is titled "On the idea of a critique of practical reason." Theoretical reason concerns "objects of the cognitive faculty only," while practical reason concerns the will, "a faculty either of producing objects corresponding to representations or of determining itself to effect such objects (whether the physical power is sufficient or not)" (5:15). Recall from the previous chapter that Kant defines the "faculty of desire," which in rational agents takes the form of practical reason, as the capacity "to be by means of [one's] representations the cause of the reality of the objects of those representations" (5:9n). Likewise, in the B-edition Preface to the first *Critique*, Kant asserts that "cognition can relate to its object in either of two ways": theoretical reason "merely determin[es] the object," whereas practical reason "mak[es] the object actual" (KrV Bix-x).[22] The *Critique of Pure Reason,* aided by reflection (in Kant's sense) on the capacities operative in theoretical cognition, demonstrated the impossibility, for discursive subjects such as ourselves, of theoretical knowledge of noumenal reality; our capacity for theoretical knowledge

[22] The distinction between theoretical and practical reason is complicated by Kant's transcendental idealism, and by the fact that, within the first *Critique*, Kant identifies a "productive" (sometimes "figurative") synthesis of the imagination that plays a role in the constitution of appearances and which ensures that appearances can be represented by categories of the understanding in acts of judgment (KrV A115ff., B150ff.). The vexed relationship of Kant's transcendental idealism to his empirical realism is beyond the scope of this dissertation. That said, it is possible that when Kant says that objects of theoretical reason are not "made," he is speaking relative to what Beatrice Longuenesse calls "the internalization, within representation, of the relation between representation and its object" (*Kant and the Capacity to Judge*, p. 20).

is thus limited to appearances, objects of empirical intuition (KrV A20/B34). Theoretical reason, then, is the capacity to represent and attain knowledge of objects given empirically.

Following Kant, it is customary to express the difference between the theoretical and practical uses of reason in terms of a difference in "standpoints" on the world.[23] The theoretical standpoint draws on sensibility and its own conceptual powers to represent given objects. By contrast, the practical standpoint draws on sensible sources of motivation (in the case of empirical practical reason) and its own pure principles to determine itself to action and thus to produce objects in accordance with its maxims. Of course, these standpoints are not mutually exclusive: any action will at the very least involve an understanding of one's immediate environment that is "theoretical" in Kant's sense, and may draw on more remote theoretical representations; likewise, theoretical reasoning itself involves undertaking to perform certain actions. Nevertheless, the difference between these standpoints is articulated in terms of their formal relationship to objects conceived from the perspective of the theoretical or practical reasoner *as objects*.

This account of the distinction between theoretical and practical reason goes some way toward refuting the claim that reason's self-critique is a *theoretical* undertaking. That is because critique is neither an attempt to know and understand the objects of the natural world, nor an attempt to produce objects in accordance with one's representation of them. In particular, a critique of practical reason, as with any critique, is not an exercise of reason that is individuated, as the kind of exercise of reason that it is, by its relationship to objects conceived as such. For a critique of reason is an essentially *reflexive* enterprise and thus concerned with the constitution of the subject conceived *as subject*—or, in the case of a critique of practical reason, as a *willing agent*. Inasmuch as critique is an essentially reflexive enterprise, to regard oneself as a willing agent is not to regard oneself as if one were an object distinct from one's rational powers, of which one can predicate various properties and capacities (for example,

[23] Cf. G 450.

the capacity for transcendentally free agency); in the relevant sense, it is not to regard oneself as an *object* at all. What this means is that a critique of practical reason is not, in the first place, an exercise of theoretical reason. And given this understanding of critique, this means that a critique of theoretical reason is likewise not an exercise of theoretical reason, but again a mode of rational examination with a distinctively reflexive character.

We can therefore make a conceptual distinction between, on the one hand, the theoretical and practical "standpoints," which are individuated by their formal relationship to objects, and, on the other hand, a *critical* perspective, which is achieved by the subject's reflecting on her own cognitive capacities in order to achieve the kind of self-cognition distinctive of critique.[24] The critical perspective has its basis in modes of apperception that are already operative *within the theoretical and practical standpoints*. To wit, our awareness of ourselves as either cognizing subjects or willing agents is made possible by the pure apperceptive character of cognizing and willing. This entails that the very distinction we make between the theoretical and practical standpoints is made *from the critical perspective*, a perspective from which we can survey our various rational capacities and articulate their roles, principles, and interconnections within our overall cognitive system. Thus, it is from the critical perspective that we distinguish between theoretical and practical reason, identify their respective principles and subject matter, and ascertain the boundaries of their legitimate use. That is, it is reason in its *critical* use, i.e., in its mode as a reflective self-examiner, that determines the scope and legitimacy of theoretical and practical reason.

Adopting the critical perspective therefore involves the presumption of a capacity for reflective self-cognition. In chapter two, I traced this presumption to one of Kant's foundational normative commitments: the assumption that reason is autonomous, and thus determines the

[24] I introduce the term 'critical perspective' (and not 'critical standpoint') to forestall the misunderstanding that in critique we adopt yet another relationship to objects conceived as such.

legitimacy of its own principles according to its own standards. I argued that the capacity for reflective self-cognition is not merely an "epistemic enabling condition" for critique's operation, but tied to critique "in conception," as an essential component of critique insofar as critique presupposes and exhibits the autonomy of reason. This is because, in order for critique to constitute an autonomous exercise of reason, the critical subject must be able to recognize her principles *as her principles* in the pure apperceptive sense. Again, this is for two reasons. First, the principles that pure reason deploys in critique must be such as to be recognized by the subject as principles she actively deploys in thought, where such thought constitutively identifies the subject as the agent of the relevant activity; otherwise, she would stand to such principles as she would an object that she represents as distinct from her own thought. Second, the principles that are submitted to critique must likewise be such as to be recognized by the subject as ones she *could* actively deploy in thought, and for the same reason. Together, these features of critique are necessary for the possibility of the subject recognizing the fulfillment of the critical task as one in which pure reason supplies the standard for its own critique.[25]

Given the foregoing account of the distinction between theoretical and practical reason, this implies that critique stands opposed to what I referred to above as a "realist" conception of justification, one in which the self-cognition of critique is a mode of theoretical reason, and the validity of critique consists in the degree to which it corresponds with an external, objective reality to which it is normatively answerable. In the Introduction to the second *Critique*, Kant invokes the conception of critique as a self-legislative enterprise in order to explain why a critique of pure practical reason—and, by implication, a deduction of the moral law—is not necessary:

> [I]f we can now discover grounds for proving that [positive freedom] does in fact belong to the human will…then it will not only be shown that pure reason can be practical but that it alone, and not reason empirically limited, is unconditionally practical. Consequently, we shall not have to do a critique of *pure practical* reason but only of practical reason as such. For, pure reason, once it is shown to exist, needs no

[25] See section 2.3.

> critique. It is pure reason that itself contains the standard for the critical examination of every use of it. (KpV 5:15-16)

When Kant claims that "it is pure reason that itself contains the standard for the critical examination of every use of it" (ibid.), he is articulating the specific method of the second *Critique* as well as speaking to a more general methodological commitment.

Specifically, the Formula of Universal Law serves as a standard for the "critique" of empirical practical reason, but that does not entail that such a critique is itself an exercise of practical reason whose internal principle is the Formula of Universal Law. Rather, consciousness of our capacity for moral thought and action constitutes a "fact of reason" that provides a sufficient basis for regarding ourselves as autonomous rational agents who face moral obligations. This in turn serves as the basis for a "critique" of empirical practical reason in the negative sense of establishing the boundaries of its legitimate deployment. Kant argues that empirical practical reason must meet a condition of universalizability, the standard of which is provided by the Formula of Universal Law. In Kant's idiom, empirical practical reason is valid insofar as it meets a condition of "rational self-love," but it cannot break out into "self-conceit," which presumes the unconditional practical authority of one's subjective inclinations (KpV 5:73).[26]

More generally, however, it is pure reason in its critical role, and not reason in its delimited, theoretical role, that determines the legitimacy of its principles. From the critical perspective, we do not presume that theoretical reason has a special authority with respect to the validity of practical reason: that would be to operate under the "realist" conception of justification that stands opposed

[26] Acting out of respect for the moral law requires us to ignore or set aside certain incentives to action, and this can often be a very painful and demanding process. Kant says relatively little about this in the *Groundwork*. It is given fuller treatment in the second *Critique*, where in the chapter titled "On the incentives of pure practical reason" Kant distinguishes self-love from self-conceit. The former is the ordinary condition of sensibly-affected, rational beings such as ourselves. When on the basis of self-love an agent ignores the ends of other rational agents, it becomes self-conceit. Kant asserts that "the moral law, which alone is truly objective…excludes altogether the influence of self-love on the supreme practical principle and infringes without end upon self-conceit, which prescribes as laws the subjective conditions of self-love" (KpV 5:74). The effect of the moral law on self-conceit is experienced as a kind of "humiliation" (5:78).

to the idea of critique as a normatively self-legislative procedure. Rather, through reflection on our theoretical and practical capacities we recognize that the theoretical and practical uses of reason have different functions within our overall rational economy. Thus, from the critical perspective, to regard the fact that we lack a theoretical proof of transcendental freedom as presenting a special problem for the legitimacy of pure practical reason would involve the arbitrary privileging of the theoretical use of reason over the practical. Contra Ameriks, who regarded Kant's failure to recognize some such problem as a lapse into dogmatism, Kant would have demonstrated a much greater departure from the critical method by arbitrarily invoking theoretical reason as the standard by which practical reason's legitimacy is to be determined. Each use of reason, within its domain, possesses a legitimacy that is freestanding with respect to the other. It is one of the tasks of critique to determine their respective domains, but critique is not itself an exercise of theoretical reason. Indeed, its very exercise is predicated on a capacity for reflective self-cognition that, on my reading of Kant, we can presume despite a lack of theoretical grounding.

A natural worry is that by eschewing a realist conception of justification, Kantian critique completely collapses any distinction between how things merely seem to us to be and how things truly are. This makes the question posed at the beginning of this section all the more pressing. Granted that the presumption of a capacity for reflective self-cognition is part of reason's presumption of its own autonomy, why didn't—and, by his own lights, why *couldn't*—Kant declare it a "fact of reason" that every event has a cause? What is distinctive about pure practical reason, qua *practical*, that the "proof" of its reality does not require a deduction? To address these questions, we must clarify the nature of the self-cognition that constitutes the basis for critique.

First, as I explained in chapter two, Kant does not assume that the mind is in all respects transparent to itself. Rather, our capacity for reflective self-cognition is grounded in the pure apperception of distinctively *rational* capacities, i.e., those capacities we engage in *actively*, according to

an internal principle that individuates that capacity as the kind of capacity that it is. Our pure apperception consists in our deploying those principles in thought (whether theoretical or practical) in such a way that (i) we are cognizant of the "function" of the relevant principle; and (ii) we are aware of ourselves as the agent "performing," or engaged in, the relevant function. In the case at hand: pure practical thought and action is attended by our pure apperceptive consciousness of determining ourselves to action on the basis of our representation that certain actions are morally required or prohibited.[27]

Nevertheless, since Kant holds that pure apperception is essential to every rational capacity *qua* rational capacity, this does amount to a restricted transparency thesis, which we can express as follows:

> A rational subject is aware, through pure apperception, of the function of the *internal principle* of the rational activity, whether theoretical or practical, in which she is engaged.[28]

By emphasizing the "internal principle" of a given rational activity, I want to draw your attention to the Kantian thesis that certain rational activities involve the cooperation of sensibility, and apply to objects of our senses. For example, consider the application of the principle that empirical objects stand in relations of cause and effect. We can characterize this activity in a way that makes essential reference to the objects to which it applies—that is, we can provide an "object dependent" characterization of its nature—so long as in doing so, we're careful to note that the pure apperception of this activity does not suffice to establish that empirical objects *do in fact* stand in relations of cause

[27] Again, this does not entail that every moral agent will possess a sophisticated grasp of the universality requirement. One of the tasks of a critique of practical reason is to elucidate the nature of pure practical reason through a precise statement of its fundamental principle.

[28] See chapter two, footnote 41.

of effect. More generally, the principles of the understanding make determinate claims about the empirical world, not claims about the internal constitution of the understanding. I am arguing, of course, that reflection on the latter is essential to establishing the former: reflection enables critique precisely by giving us, the critical subjects, an epistemic foothold into the nature of our own cognition. But claims about the nature of objects given in sensibility lie outside the purview of the kind of reflective self-cognition that Kant takes to constitute the basis for—rather than the achievement of—critique. *This* is why Kant could not postulate, in advance of critique, that every event has a cause.

The above distinction coincides with a significant difference between Kant's critiques of theoretical and practical reason, and Kant's sensitivity to this distinction is apparent in the Preface and Introduction to the second *Critique*. First, consider Kant's announcement in the Preface that "if as pure reason it is really practical, it proves its reality and that of its concepts" *durch die Tat* (KpV 5:3). In light of the foregoing remarks, we can now explain what is distinctive about pure practical reason, *as practical*, such that it does not require its own critique. A moral agent's pure apperception of her pure practical thought and action constitutively identifies that agent as the agent of the relevant activity, that is, as someone capable of determining her will according to her representation of some possible action as morally required or prohibited. This capacity is the practical species of the more general capacity for pure apperception described above. In the paradigmatic case, practical reason produces an object through its representation. But the pure apperception of pure practical reason does not essentially involve the representation of a given object according to an *a priori* concept or principle. In the idiom of the restricted transparency thesis set forth above, apperception of the function of the "internal principle" of one's activity suffices for the reflective cognition of oneself as a moral agent. And since such apperception is a constitutive feature of that very activity, there is a straightforward sense in which the capacity for pure practical reason is "proved" through its very operation—proved, that is, *durch die Tat*. Of course, the same cannot be said for the principles of the understanding. Pure

apperception of the function of those principles does not by itself establish, e.g., that empirical objects actually stand in relations of cause and effect.

In the Introduction that follows, Kant invokes this distinction to explain why a critique of practical reason should begin with an articulation of the "principles" of practical reason (5:16).[29] He does this immediately after claiming that "pure reason…contains the standard for the critical examination of every use of it," and thus that pure practical reason provides the basis for a critique of "empirically conditioned" practical reason (ibid.). Because a critique of practical reason is concerned with the determination of the will[30] according to principles and not the application of *a priori* principles to empirically given objects, such a critique need not proceed from an account of our sensibility:

> [T]he order in the subdivision of the Analytic [of the *Critique of Practical Reason*] will be the reverse of that in the *Critique* of pure speculative reason. For, in the present *Critique* we shall begin with *principles* and proceed to *concepts*, and only then, where possible, from them to the senses, whereas in the case of speculative reason we had to begin with the senses and end with principles. The ground for doing so lies, again, in this: that now we have to do with a will and have to consider reason not in its relation to objects but in relation to this will and its causality; thus the principles of empirically unconditioned causality must come first, and only afterward can the attempt be made to establish our concepts of the determining ground of such a will, of their application to objects and finally to the subject and its sensibility. Here the law of causality from freedom, that is, some pure practical rational principle [i.e., the Formula of Universal Law], constitutes the unavoidable beginning and determines the objects to which alone it can be referred. (5:16, my emphasis)

Kant's proposed outline reflects the three-chapter structure of the Analytic of the second *Critique*. We have focused on certain of Kant's prefatory remarks concerning method, as well as aspects of Chapter One ("On the Principles of Pure Practical Reason"). Chapter Two addresses "the concept of an object

---

[29] I believe Kant refers to "principles" and not the one supreme principle of pure practical reason because he also has in mind the restriction of principles of empirical practical reason to conformity with the former, as well as the broader system of categorical and hypothetical imperatives. In the same passage he refers to "principles of empirically unconditioned causality," by which he means the various categorical imperatives that constitute our moral duties (5:16).

[30] I use the term 'determination of the will' because it is Kant's own and because it includes the case where, due to physical incapacity, no outward behavior is manifested. See once again KpV 5:15, where Kant describes practical reason as reason's capacity to "[determine] itself to effect…objects (whether the physical power is sufficient or not)." As I argue in the previous chapter, we can think of this as a limiting case of action.

of pure practical reason," arguing for the priority of the moral law over our conception of the good (5:57ff.). Finally, Chapter Three concerns how the moral law can act as an "incentive" for sensibly affected agents such as ourselves (5:71ff.). By contrast, the Analytic of the first *Critique* established the transcendental conditions of the application of *a priori* principles to objects given in space and time, as well as the restriction of the principles to such objects in their legitimate use, and thus the impossibility for discursive subjects of knowledge of things-in-themselves. But in order to establish this conclusion, Kant first had to establish, in the Transcendental Aesthetic, that space and time are mere forms of our sensibility.[31]

Later, in a section of Chapter One titled "On the Deduction of the Principles of Pure Practical Reason," Kant argues that a deduction of the moral law need not and, as we saw in the last chapter, cannot be given. In the course of making this argument, Kant again emphasizes the inversion of the order of exposition of the first two *Critiques*, specifically the fact that the second *Critique* proceeds from principles of the will's determination. After describing the "problem" of the first *Critique*, i.e., the problem of synthetic, *a priori* cognition of objects, Kant continues:

> The second [problem], which belongs to the *Critique of Practical Reason*, requires no explanation of how objects of the faculty of desire are possible…but only how reason can determine maxims of the will, whether this takes place only by means of empirical representations as determining grounds or whether pure reason might also be practical… Whether the causality of the will is adequate for the reality of the objects or not is left to the theoretical principles of reason to estimate, this being an investigation into the possibility of objects of volition, the intuition of which is accordingly no component of the practical problem. …
>
> In this undertaking the *Critique* can therefore not be censured for beginning with pure practical laws and their reality, and it must begin there. (5:45-46)

In this passage Kant again contrasts the second *Critique* with an investigation into the possibility of the *a priori* cognition of objects. Indeed, here he claims that a critique of practical reason does not even concern itself with the question of whether the will's determination from pure rational principles is

[31] Cf. KpV 5:42.

sufficient for the intended effect, this being a problem for theoretical reason. Most importantly for our purposes, however, is the fact that Kant *explicitly* cites this feature as the reason that the *Critique of Practical Reason* may assume "pure practical laws and their reality." The foregoing remarks provide a way of spelling out this connection: in critique, we are entitled to assume a power of reflective self-cognition that discloses the function of the principles of our rational activity. Since a critique of practical reason is concerned *only* with the principles for the determination of the will, the pure apperception that attends moral agency can be regarded as a "fact of reason," one that suffices as a basis for regarding ourselves as pure practical reasoners.

Although the exposition of the Analytic of the second *Critique* inverts the order of the first *Critique*, we should still expect there to be a role within the first *Critique* for the reflective cognition of principles, where the assumption of the legitimacy of those principles does not necessitate a deduction. That is, we should expect there to be principles the articulation of which likewise constitutes the "unavoidable beginning" for further critical inquiry.[32] In chapter two, I proposed that the first *Critique*'s articulation of twelve distinct functions of judgment is grounded in reflection on the general logical use of the understanding. In this respect, the various general logical principles that are each associated with a distinct logical form are analogous to the Formula of Universal Law, in that reflective cognition of them is made the basis for critical investigation, and no "deduction" is offered in the service of proving that such principles in fact constitute the general logical form of our thinking. The structure of the Analytic of the first *Critique* suggests exactly such a parallel. Broadly, it consists of five stages: an articulation of the general logical form of our thought, a "metaphysical deduction" that associates each logical function with a specific category, or pure concept of the understanding, a "transcendental deduction" that establishes the possibility of such concepts applying to given objects, an account of the spatiotemporally "schematized" versions of these concepts, and finally, an elaboration of the

[32] Cf. KpV 5:16.

principles of the understanding that deploy these concepts and to which empirical objects necessarily conform. Within the Analytic, the logical functions of the understanding are presented as functions of the most fundamental principles of thought itself, and are invoked by Kant in an account of the transcendental conditions of empirical cognition. More specifically, within the metaphysical and transcendental deductions of the categories, the logical functions of judgment are those fundamental capacities partly in terms of which categorial representation of objects is to be understood.[33] Like the Formula of Universal Law, general logical principles and their associated functions are items of reflective cognition that are made the basis for subsequent arguments within the relevant critique. If this is right, then the second *Critique's* reliance on self-consciousness to set forth the fundamental principle of a particular rational capacity has a broad precedent in Kant's critical philosophy.[34]

This concludes the positive argument for why a deduction of the moral law is not necessary. I have argued that the doctrine of the fact of reason is grounded in the presumption of a capacity for reflective self-cognition. While this falls short of an unrestricted transparency thesis, it does entail a capacity for cognizing the nature of our own subjective, rational activity—the "function" of the principles internal to those activities. In the case of moral thought and action, we are aware, through our pure apperception of that activity, of our capacity to determine our wills on the basis of our representation of some candidate action as morally required or prohibited, and this suffices to entitle us to regard ourselves as moral agents. Moreover, this presumption is an essential feature of Kant's

[33] In saying that our cognition of the general logical form of thought is grounded in reflection, I do not mean to imply that we have unmediated access to the general logical form of our thinking apart from the role of such forms in our cognition of objects. My point is only that, because Kant holds that we have reflective access to those logical forms, there is no attempt on Kant's part to prove that such forms are in fact the logical forms of our thinking, i.e., to provide a "deduction" of the general logical form of thought. That point is compatible with the idea that we can only identify the forms of pure general logic by prescinding from the form of our cognition of objects, in which case logical reflection would be mediated by transcendental reflection. Again, the characterization of general logic as freestanding with respect to transcendental logic has been disputed by Longuenesse (*Kant and the Capacity to Judge*, p. 76). See chapter two, footnote 36.

[34] This is in addition to the earlier noted parallel between the doctrine of pure apperception in the Transcendental Deduction and the doctrine of the fact of reason in the second *Critique*. I do not mean to suggest any exact analogies, but instead broad parallels that demonstrate a continuity of method across the first two *Critiques*.

eschewing a "realist" conception of justification for one according to which pure reason provides the standard for its own self-critique. If, as I argued above, theoretical reason does not supply the standard of practical reason's legitimacy, Kant's rejection of this demand is perfectly consistent with critical methods.

Nonetheless, even if one grants that the doctrine of the fact of reason does not constitute a regression into dogmatism, one might worry that Kant's practical philosophy is incompatible with his theoretical philosophy. For Kant's practical philosophy appears to claim a kind of cognition of which his theoretical philosophy denies the very possibility, namely, cognition of our transcendental freedom.[35] This brings us to the conclusion of this section.

In the Preface to the second *Critique,* Kant addresses the apparent inconsistency by explicitly denying that we possess theoretical cognition of our freedom: "we cannot theoretically cognize and have insight into" our freedom (KpV 5:4); "reason is not thereby extended in theoretical cognition" (5:5); "the reality thought of here does not aim at any theoretical *determination of the categories* and extension of cognition to the supersensible" (ibid.); rather, "what is meant [by the claim that the reality of freedom is established] is only that in this respect an *object* belongs to [freedom], because [freedom is] contained in the necessary determination of the will a priori" (ibid.).[36] This is a theme that was already present in the *Groundwork*. There Kant claims that the assumption that we act "under the idea of freedom" entails that we are "really free in a practical respect" (G 4:448), but makes clear that this

[35] For clarity of exposition, I ignore the fact that Kant also claims that the ideas of God and immortality are given "objective reality" through their connection to the moral law (KpV 5:4). According to Kant, belief in the reality of freedom, God, and immortality is justified by the role that their corresponding ideas play in the operation of pure practical reason. However, freedom has a special place among these ideas. Only freedom is a "condition of the moral law" (ibid.), that is, a necessary condition of the capacity to be motivated by the representation of duty and thus be subject to the categorical demands of morality. God and immortality, by contrast, are "conditions of the necessary object of a will determined by this law," the highest good (ibid.). The role of the highest good in Kant's practical philosophy is beyond the scope of this dissertation.

[36] For consistency of exposition, I've modified the last sentence to emphasize the "practical reality" Kant claims for freedom. See prior footnote.

result does not constitute a proof of freedom "in a theoretical respect" (4:448n). According to the *Groundwork,* our entitlement to regard ourselves as free partially resides in the fact that freedom is "a necessary presupposition of reason in a being that believes itself to be conscious of a will" (4:459).[37]

Both the *Groundwork* and the second *Critique* ultimately claim that we are entitled to regard ourselves as transcendentally free beings who are genuinely morally obligated. Of course, the latter rejects the demand to establish the legitimacy of pure practical reason through a special critique of the "pure" faculty. But setting this difference aside, they share the view that our entitlement does not issue from or consist in any sort of amplification of our theoretical knowledge beyond the limits of experience. This is because the claim that we are so entitled is restricted, sometimes explicitly but always at least implicitly, to reason in its *practical* use: the ideas of freedom and moral obligation, as "necessary presuppositions" of our faculty of practical reason, have a practical but not theoretical legitimacy. Kant therefore avoids the charge of inconsistency by restricting the scope of the legitimacy of various cognitive activities and "presuppositions," including the presupposition of freedom identified in practical reflection.

But what does this restriction come to? I said earlier that Kant's doctrine of the fact of reason—his insistence that pure practical reason proves its reality by its own activity, which obviates the need for a special critique of the "pure" practical faculty—is partly to be explained by the fact that from the "critical perspective" we do not presume that the theoretical use of reason has special authority with respect to the practical use. Central to this line of thought is the claim that the theoretical

[37] Given all that is said here, one might wonder how Ameriks could have possibly thought that Kant intended the deduction of *Groundwork* III as a theoretical argument for our transcendental freedom and moral agency. On Ameriks's view, Kant believed that we could have *theoretical proof* of our transcendental freedom but no insight into its noumenal mechanics. That is, according to Ameriks, Kant thought that we could possess theoretical knowledge *that* we are transcendentally free but no such knowledge concerning how such freedom is possible ("Kant's Deduction of Freedom and Morality," p. 60). On this view, Kant's claim that we are "free in a practical respect" is understood to involve the availability of a *theoretical* proof of such freedom. I do not directly argue against this component of Ameriks's view, but I believe it requires a highly idiosyncratic and ultimately quite implausible reading of Kant's restriction of theoretical knowledge to phenomena, as well as what is meant by Kant's restriction of the assumption of our transcendental freedom to the practical use of reason.

and practical uses of reason carve out distinct cognitive domains: a use of reason is legitimate so long as it does not "go…beyond its sphere" (KpV 5:16).[38] A natural worry is that this amounts to little more than definitional fiat, with Kant labeling uses of reason either 'theoretical' or 'practical' only in order to avoid explicit contradiction.[39] In response to this worry, I will try to explain in greater detail how Kant conceives of this distinction.

To begin, let's return to how Kant distinguished the theoretical and practical uses of reason in the Introduction to the second *Critique*. That distinction was explained in terms of a formal difference in the relationship of a representation to its object. Whereas theoretical reason represents given objects, practical reason *produces* its objects; again, practical reason is the rational species of the faculty of desire, the capacity "to be by means of [one's] representations the cause of the reality of the objects of those representations" (KpV 5:9n). We can articulate this difference in terms of different cognitive tasks or aims: it is the *aim* of an exercise of theoretical reason to conform its representations to given objects, and it is the *aim* of an exercise of practical reason to realize the objects of its representations. Furthermore, these are *constitutive* aims, aims that constitute a given exercise of reason as the kind of exercise of reason that it is.

Thus, in the paradigm cases, an exercise of theoretical reason culminates in a representation (a belief or judgment) and an exercise of practical reason culminates in an action, understood here as a determination of the will; and this difference owes to the fact that these exercises have different constitutive aims. One way of drawing out this difference is by examining how an apparently identical

---

[38] In the quoted passage, Kant refers to the sphere of the empirically conditioned use of practical reason, but the metaphor of a "sphere" as describing a domain of legitimate use can be generalized to all rational faculties.

[39] No doubt invoking this worry, Mahrad Almotahari once characterized one of our disagreements as follows:

> M.A.: From the theoretical perspective, *p*; from the practical perspective, *not-p*; problem!
> J.B.: What do you mean? From the *theoretical* perspective, p; from the *practical* perspective, not-p; no problem!

question can be posed in either a "theoretical" or "practical" voice, that is, with these different constitutive aims in mind.

Consider a question of the form "What should I do?" When asked in a practical voice, any deliberation addressed to this question will be such as to determine one or several possible courses of action. This is so even if, owing perhaps to physical incapacity or weakness of will, I do not do as I judge (practically) that I ought. But insofar as I take myself to be responding to a practical question, my deliberation regarding what I *should* do will have immediate relevance to the question regarding what *to* do.

Suppose, however, that I ask the question "What should I do?" in a *theoretical* voice. Here I am not asking this question with a view to deciding what to do. Rather, I am inquiring about objects or features of the world, broadly understood, that might be taken to bear on the factual question of what I should do. For example, perhaps I am asking whether there exists a reason R to perform some action A, with a view to (i) believing that R exists should I judge that there is sufficient evidence that R exists, (ii) believing that R does not exist should I judge that there is sufficient evidence that R does not exist, or (iii) remaining committedly neutral on the question of R's existence should I judge that the evidence is inconclusive.[40] Or perhaps I am taking a class on ontology and speculating about the ontological status of reasons to act, in particular whether there exist such reasons.[41] What distinguishes these as exercises of theoretical reason is that when, as the outcome of my deliberation, I judge, say, that R exists, I do not take this to have immediate practical relevance. My theoretical judgment that a reason

[40] For a contemporary view along these lines, i.e., one that conceives of practical reasoning as a *theoretical* task, see Scanlon (*Being Realistic about Reasons*).

[41] In such a case, I might be asking (again, in a theoretical voice) "What, *if anything*, should I do?" Here I assume that such a question *can* be posed in a theoretical voice. The fact that it seems unnatural to frame ontological inquiry as a response to such a question is, I believe, an illustration of my point. Because ontological inquiry doesn't have direct bearing on how I should determine my will, it seems odd to characterize such inquiry in terms of the question of what I should do. Kantians, of course, believe that a "theoretical" account of practical reason mischaracterizes the subject matter at a fundamental level. And Kant would deny that we could have theoretical knowledge of this sort. My point in framing these questions in this way is just to draw out the contrast between theoretical and practical reasoning as distinct cognitive tasks.

R exists for me to perform some action A has no immediate bearing on the question of what to do, even though such a representation constitutes an answer—but only in a theoretical sense—to the question of what I should do. Likewise, my seminar room conviction that there are tables and chairs and electrons but nothing quite so mysterious as *reasons* has no immediate bearing on what to do. This is because theoretical reasoning is not essentially such as to determine action.

In short, what distinguishes the uses to which we can put this question is the *cognitive task* involved in answering it. Thinking of exercises of theoretical and practical reason as constituted by different cognitive aims is crucial to understanding how the idea of freedom can have self-standing practical legitimacy and why the "objective though only practical reality" afforded to this idea (KpV 5:48) does not involve any violation of the restriction of theoretical knowledge to appearances. The basis for regarding ourselves as autonomous agents who are morally obligated is the apperception of pure practical deliberation and action. To say that the ideas of freedom and moral obligation have practical legitimacy is to say that they are warranted presuppositions *of the activity of practical reasoning*. To be sure, Kant is not saying just that we cannot, as a brute psychological matter, help but think of ourselves as moral agents when we act; it's rather that we are justified in doing so. However, this justification is restricted to the task at hand: we are specifically entitled to regard ourselves as moral agents *for the purpose of* engaging in the cognitive task of reasoning about what to do.[42]

Because our entitlement to regard ourselves as autonomous moral agents is relative to a particular cognitive aim, the doctrine of the fact of reason should not be taken to supplant theoretical doubts we might raise about freedom and moral obligation. The fact that I am entitled to regard myself as a moral agent for the purposes of practical reasoning does not entail that I am entitled to this assumption for the purposes of theoretical reasoning. (Indeed, Kant affirms that we are *not* entitled to

[42] Note that this is not the same as saying that we are entitled to regard ourselves this way *when* we engage in that cognitive task. I have formulated the point so as to make explicit that our entitlement is relative to a particular cognitive aim.

this assumption, since theoretical cognition is confined to appearances.) This helps to explain why the "practical reality" afforded to freedom does not involve an extension of theoretical knowledge beyond what is given empirically. For Kant, our entitlement to regard ourselves as moral agents has no theoretical significance whatsoever.

4.4 I turn now to the deduction of morality that Kant attempts in the third chapter of the *Groundwork of the Metaphysics of Morals* ('*Groundwork* III'). This is for two reasons. First, recall that according to Ameriks, Kant conceives of the deduction he attempts in *Groundwork* III as a *theoretical* argument for the existence of moral obligation. If that is right, it spells trouble for my contention that philosophical critique is not an exercise of theoretical reason, for Kant subsumes the argument of *Groundwork* III under the banner of a "critique of pure practical reason" (G 4:446). In response, I argue that while Kant carefully avoids appealing to *moral* considerations in his attempted deduction—even worrying about an implicit circle in the inference from freedom to morality—it is extremely doubtful that the premises of that argument are, or were intended to be, "theoretical" in Kant's sense. Rather, *Groundwork* III is addressed to beings who take themselves to be practical reasoners, and it avails itself of considerations made available through practical reflection. In particular, Kant appeals to practical reflection both in the assumption that we are practical reasoners, as well as in the claim that, as practical reasoners, we necessarily act "under the idea of freedom."

Second, the argument of *Groundwork* III is instructive in the way that it relies on items of self-cognition made available through reflection on our practical capacities, and yet treats the fact of the "interest" we take in morality as something that needs to be *proved*. I will argue that there is a certain tension implicit in the combination of this argument's methods and aims, and this bears on the question of whether a deduction of the moral law *needs* to be given, that is, whether our rational entitlement to regard ourselves as agents who face moral obligations depends on the availability of

such an argument. To that end, I will not offer a complete exegesis of *Groundwork* III, but instead focus on those aspects of the chapter that are directly relevant to the above concerns.[43]

Groundwork III begins by introducing some statements and terms that are by now familiar. The will, which in *Groundwork* II was identified with practical reason (G 4:412),[44] is said to be "a kind of causality of living beings insofar as they are rational" (4:446). Negative freedom is defined as "that property of [the will] that it can be efficient independently of alien causes *determining* it," and Kant asserts that a "positive" notion of freedom "flows from" the former (ibid.). Our capacity for positive freedom is, of course, autonomy, our power to determine ourselves to action on the basis of principles of pure reason, here described as "the will's property of being a law to itself" (4:447). Finally, Kant argues that the principle of such autonomy is the Formula of Universal Law, concluding on that basis that "a free will and a will under moral laws are one and the same" (ibid.).

But as we noted in the last chapter, it is not sufficient for Kant's purposes merely to establish an analytic connection between transcendental freedom and the authority of the moral law. For it is neither an analytic truth nor an empirically demonstrable statement that rational beings such as ourselves are transcendentally free. At this stage of the argument, Kant is concerned to show that morality is not a "phantom of the human imagination" (G 4:407) or "chimerical idea without any truth" (4:445). To establish the synthetic, *a priori* claim that rational human beings are morally obligated, Kant sets out to provide a "deduction" of morality, the positive task of a putative critique of *pure* practical reason.

---

[43] For example, I ignore Kant's argument for the possibility of categorical obligation, which he provides in the section titled "How Is A Categorical Imperative Possible?" Kant offers this argument after he takes himself to have shown a non-moral basis for the "interest" that we take in morality. Kant's aim in this section is to explain why, given our interest in morality, we represent pure practical principles as duties or imperatives. See G 4:453-455. Moreover, my treatment of the "deduction of freedom" (the argument Kant provides to overcome a covert "circle" he alleges to find in the preliminary argument of *Groundwork* III) will be cursory.

[44] See chapter three, footnote 18.

It is worth taking a moment to consider the motivation behind this task, that is, the sorts of considerations that seemed to Kant at the time to necessitate a non-circular argument for the existence of moral obligation. For as I mentioned in the last chapter, Kant does not seriously doubt that we are morally obligated. In a statement reminiscent of the second *Critique*'s doctrine of the fact of reason, he even asserts, after setting forth his deduction, that "the practical use of common human reason confirms" the results of the argument he has just provided (4:454, my emphasis). Furthermore, as I shall argue, Kant already held, at the time of the writing of the *Groundwork*, the pure apperceptive account of moral consciousness that I've claimed is put at the basis of the second *Critique* as the doctrine of the fact of reason. Assuming this is right, it raises the question of why the *Groundwork* doesn't follow the path of the second *Critique*. In other words, why doesn't Kant in the *Groundwork* set forth such consciousness as the "ratio cognoscendi" of our transcendental freedom and the basis for a critique of practical reason?

Part of the answer consists in the fact that we can, when adopting the theoretical standpoint on the world, raise genuine doubts about both the possibility of categorical obligation as well as our capacity for pure practical reason. In *Groundwork* II, Kant makes it a point to emphasize that we cannot achieve empirical cognition concerning whether we have in fact acted on the basis of our representation of duty (4:407). While we sometimes regard ourselves, from the practical standpoint, as having acted from the representation that some action is morally required, from the theoretical standpoint we cannot know that a "covert impulse of self-love, under the mere pretense of [duty], was not actually the real determining cause of the will" (ibid.). Moreover, while we indeed recognize the authority of moral standards whenever we adopt the practical standpoint, we can nevertheless interrogate what—if anything—might be the source of the legitimacy of those standards. The basis of such "doubt" consists in the fact that we can take a sidelong glance at our own practical deliberation, even entertaining the possibility that morality might be a "chimerical idea." But as moral agents, this

is a possibility that is in a certain sense closed off to us in deliberating about what to do. Thus, the *Groundwork* attempts to legitimize what is apparent to us *within* the practical standpoint, from a perspective alienated *from* that very standpoint.[45]

The deduction begins with a preliminary argument presented in a section titled "Freedom Must Be Presupposed as a Property of the Will of All Rational Beings" (4:447). There Kant puts forth two major claims:

(1) "[E]very being that cannot act except under the idea of freedom is just because of that really free in a practical respect," and consequently, "that all laws that are inseparably bound up with freedom hold for him just as if his will had been validly pronounced free also in itself and in theoretical philosophy" (4:449). The ensuing discussion makes clear that Kant means at least the following: a rational agent who, whenever she acts, does so "under the idea of freedom" is one who presupposes her own transcendental freedom in practical deliberation. That is, whether or not she is convinced, as a theoretical matter, that her actions are determined by external causes, whenever she engages in practical deliberation she must do so as if she had the capacity for transcendentally free choice. Therefore, according to Kant, she must regard as authoritative the principle of such freedom, the Formula of Universal Law.[46, 47]

---

[45] However, *Groundwork* III does not disavow the practical standpoint entirely. As I will argue, the argument freely avails itself of items of self-cognition made available through reflection on our practical capacities. So while the argument is motivated by considerations that have their basis in theoretical reason, the argument is not itself theoretical in Kant's sense.

[46] Cf. Korsgaard ("Morality as Freedom," pp. 162-163). The above gloss ignores certain interpretive issues introduced by Kant's invoking an *idea* of freedom. In "The Form of Autonomy and the Deduction of Freedom in *Groundwork* III" (unpublished), I argue that the role of the idea of freedom is analogous to the regulative function of the ideas of pure theoretical reason, which Kant introduces in the Dialectic of the first *Critique*. David Sussman has a similar reading of this section of the *Groundwork* ("From Deduction to Deed: Kant's Grounding of the Moral Law," p. 56). Sussman calls our attention to the discussion of practical ideas in the second *Critique*. There Kant claims that "the moral ideas, as archetypes of practical perfection, serve as the indispensable rule of moral conduct and also as the *standard of comparison*" (KpV 5:127n). Such issues take us beyond the scope of this dissertation.

[47] Notice also that this account of what it means to act "under the idea of freedom" elides the distinction between negative and positive freedom. This will become relevant when we introduce Kant's worry about a covert circle in the preliminary argument.

(2) "[T]o every rational being having a will we must necessarily lend the idea of freedom also, under which alone he acts" (ibid.).

According to (1), if an agent cannot act except under the idea of freedom, then she is "really free in a practical respect" and thus bound by the Formula of Universal Law. But according to (2), the antecedent holds for every rational being, i.e., *every* rational agent necessarily acts under the idea of freedom. Kant's argument for the second claim is the following:

> In such a being [i.e., a rational agent] we think of a reason that is practical, that is, has causality with respect to its objects. Now, one cannot possibly think of a reason that would consciously receive direction from any other quarter with respect to its judgments, since the subject would then attribute the determination of his judgment not to his reason but to an impulse. Reason must regard itself as the author of its principles independently of alien influences; consequently, as practical reason or as the will of a rational being it must be regarded of itself as free… (G 4:448)

This passage was cited in chapter two as evidence for the claim that, according to Kant, reason necessarily presupposes its own autonomy. On this view, exercises of reason, whether theoretical or practical, can never self-consciously yield judgments on grounds other than reason's own principles. In the immediate context of the passage, Kant is claiming that a rational agent cannot consciously surrender the authority of her own reason to outside influences. To suspend or subordinate the judgment of one's own reason to an external factor—such as another's judgment, or an external incentive—is not to regard one's judgment as justified on the basis of one's own rational reflection, but rather as determined from without. This is as true for theoretical reason as it is for practical reason. Judgment, whether theoretical or practical, is the outcome of a rational activity that employs various normative standards, and the subject or agent regards her judgments as justified precisely because they conform to those standards. But if one conceives of one's judgment as caused or compelled by factors other than one's own deliberation, then one does not take one's judgment to conform to those normative standards, or at least does not take such conformity to be a necessary precondition of one's having formed the judgment one did. Consequently, the judgment is not regarded by the agent or

subject as the necessary outcome of an activity of deliberation and thus is not regarded as justified. Moreover, since one necessarily *does* take one's judgments to be justified, that means that one does not regard oneself as having formed a *judgment* at all. This is the sense in which practical reason must "regard itself as the author of its principles independently of alien influences" and consequently as free (4:448).

Let us pause to consider the significance of this argument for our understanding of the greater "deduction" that encompasses it. Even if we grant that we are rational agents (i.e., practical reasoners, or beings who possess wills), it is by Kant's lights a synthetic, *a priori* claim that any or all rational agents presuppose their transcendental freedom when they act: the concept of transcendental freedom is not analytically contained in the concept of rational agency. In chapter two, I claimed that reflection on our cognitive capacities enables critique by providing it with a foundation of *substantive* claims about the nature of our own cognition. With this in mind, note that in the passage cited above, Kant appeals to substantive, essential features of our capacity for judgment precisely in order to justify a synthetic, *a priori* principle, namely, that all rational agents act under the idea of freedom. For this reason, I want to suggest that the above passage articulates *reflective* self-cognition, that is, items of self-cognition that have their basis in the pure apperception of our own rational activity. Note in particular that Kant defends the idea that reason constitutively presupposes its own freedom by pointing to the absurdity of a subject consciously "attribut[ing] the determination of his judgment…to an impulse" (ibid.). Contra Ameriks, the impossibility of this scenario could not have a theoretical justification, for we are not considering rational agents as appearances but rather in terms of the inner constitution of their rational agency. What, then, could be the justification for such a claim? On my view, Kant is presenting us with a claim concerning the nature of rational subjectivity of *precisely* the sort that has its epistemological basis in reflection. That is, Kant is expecting us to assent to this principle not because he believes it to have a theoretical justification, but because he holds that reflection enables us to

identify the character and presuppositions of our own capacity for judgment. More specifically, he is implicitly inviting us to reflect on our capacity for practical judgment and, through consideration of the absurdity of consciously determining such judgment according to the dictates of an external influence, identify *in our own capacity for practical judgment* a presupposition of transcendental freedom.

I believe that Kant likewise appeals to practical reflection in the very assumption that we are rational agents. But to motivate this claim, some stage setting is required. The obvious though unstated implication of the preliminary argument is that because we are rational agents, we are likewise bound by the moral law. But instead of developing the argument in this natural way, Kant unexpectedly pivots to the topic of the "interest" that we take in morality, and asserts that the preceding argument contains "a kind of circle": that argument tried to demonstrate that we regard ourselves as subject to moral laws *because* we regard ourselves as transcendentally free; however, according to Kant, we only assumed our freedom "in order to think ourselves under moral laws" (4:450). Consequently, Kant sets for himself the task of explaining or justifying the "interest" we take in morality on grounds that do not presuppose the authority of the moral law. Kant defines an "interest" as "that by which reason becomes practical, i.e. becomes a cause determining the will" (4:460n). Thus, Kant is inquiring into how it is that we are motivated to obey the moral law—how it is that our pure "respect" for the moral law is a motivating consideration.

Perhaps the greatest challenge for interpreters of this section is that it is extremely unclear what "circle" Kant is alluding to. Nowhere in the preliminary argument does Kant argue from explicitly moral premises, and for that reason it can seem as if Kant misrepresents his own argument. At the same time, Kant's worry about a circle is not exactly a concession that the preceding argument *fails*. Kant introduces that argument as "preparation" for subsequent exposition (4:447). He even begins the section in which he addresses the possibility of a circle with a restatement of the conclusion

of the preceding one: "we must presuppose [freedom] if we want to think of a being as rational and endowed with consciousness of his causality with respect to actions, that is, with a will" (4:449).

The conclusion of the preliminary argument is a conditional statement: *if* a being is a rational agent, then that agent must act under the idea of freedom. It follows that if we regard ourselves as rational agents, we must likewise act under the idea of freedom. Some commentators have argued that the "circle" of the preliminary argument owes precisely to the fact that it applies to rational agents. On this interpretation, there is nothing *per se* circular about the argument, but it is question begging to assume that we are rational agents and thus that the preliminary argument applies to us.[48] But there are compelling—indeed, to my mind, decisive—reasons to reject this interpretation.[49] If that is right,

[48] Such commentators include Henry Allison (*Kant's Theory of Freedom*) and Dieter Henrich ("The Deduction of the Moral Law: The Reasons for the Obscurity of the Final Section of Kant's *Groundwork of the Metaphysics of Morals*").

[49] According to Henry Allison, given the conditional form of the conclusion of the preliminary argument, to assume that we are rational agents without providing further support for this claim is to presume our own freedom in order to regard ourselves as morally obligated (*Kant's Theory of Freedom*, p. 221). But there are several problems with this interpretation. First, Allison does not motivate the claim that the assumption that we are rational agents implicitly assumes that we are bound by the moral law. In order for Allison's interpretation to work, he would have to explain why Kant does not have the resources to defend the claim that we are rational agents except on moral grounds. On my interpretation, Kant's assumption that we possess wills is grounded in practical reflection. Second, Kant's avowed reason for worrying about a circle is the "reciprocal" connection between morality and freedom; since the preliminary argument implicitly relies on this connection, this implies that Kant's worry would remain in place even if he had provided an additional argument validating the assumption that we are rational agents. Finally, and most damning of all for Allison's interpretation, Kant's proposed solution to circumventing the alleged circle itself relies on the assumption that we regard ourselves as possessing a will. A central claim of that "deduction of freedom," which Kant takes to succeed in removing the suspicion of a circle, is not that we possess a will, as Allison supposes, but rather that "as a being belonging to the intelligible world, the human being can never think of the causality of his own will otherwise than under the idea of freedom" (G 4:452, my emphasis). It is doubtful that this is a careless slip of the pen, for Kant includes some variant of the qualification that the argument applies to beings with (or conscious of possessing) a will at various points throughout the subsequent discussion (4:457, 4:459, and 4:461). Similarly, Dieter Henrich has argued that the premise that we are conscious of possessing a will signals Kant's implicit acknowledgment that he cannot completely avoid the circle. The most Kant can do, according to Henrich, is argue that the assumption that we are transcendentally free is not completely arbitrary, because it parallels commitments made by theoretical reason, which also makes a distinction between the sensible and the intelligible. On this view, Kant cannot give a proof or "strong" deduction of the moral law from non-moral premises, because we do not have any non-moral basis for regarding ourselves as transcendentally free ("The Deduction of the Moral Law: The Reasons for the Obscurity of the Final Section of Kant's *Groundwork of the Metaphysics of Morals*," p. 338). Henrich also suggests that the premise that we are conscious of possessing a will is made "virtually inscrutable" in the service of avoiding "at least verbally" the claim, made explicit in the second *Critique*, that we have no non-moral grounds for regarding ourselves as transcendentally free (ibid.). On this particular point, Henrich's interpretation is ambiguous between whether Kant was deliberately obscure in order to conceal the weaknesses of his deduction or just confused about the relative "strength" of the deduction he could provide. Henrich's interpretation has the virtue of purporting to show a great deal of continuity between the position of the *Groundwork* and that of the second *Critique*, but it is difficult to reconcile this interpretation with the decisive tone with which Kant unequivocally declares himself to have circumvented the worry of a circle. At the conclusion of the argument in which Kant states that we represent ourselves as autonomous members of an "intelligible world," Kant declares that

then Kant does not regard the claim that we are rational agents to be a problematic assumption for the purposes of a deduction of the moral law.

This raises the question of why Kant believed he was entitled to such an assumption, and here again I claim that Kant is making an implicit appeal to practical reflection. The aim of the deduction of *Groundwork* III is to establish, on non-moral grounds, the "right" of rational agents to regard themselves as morally obligated. But that is compatible with the argument presupposing at least a *non-moral* capacity for practical reason and addressing itself to us in our capacity as such reasoners. Given Kant's distinction between empirical and pure practical reason, that is just to say that the argument is addressed to those who are, and who take themselves to be, empirical practical reasoners, i.e., individuals who can rationally pursue various ends by means of hypothetical imperatives of prudence and skill.[50] But as before, our capacity for empirical practical reason is not an item of theoretical knowledge. Rather, we regard ourselves as empirical practical reasoners on the basis of our pure apperception of the relevant activity, in this case, the activity of setting empirically-incentivized ends and electing to act on maxims in their pursuit. But if that is right, then presumably the argument of *Groundwork* III presupposes that we are rational agents because Kant holds that our capacity for rational agency, including our capacity for non-moral agency, is an item of reflective self-cognition. This marks the second major thesis of *Groundwork* III that has its basis in practical reflection.

"the suspicion [of a hidden circle] is now removed," and that the moral law is a "demonstrable [*erweislichen*] proposition" (G 4:453). The adjective '*erweislich*' means provable or demonstrable, and is a cognate of '*Erweis*' (proof) and '*erweisen*' (to prove). This suggests that Kant sees himself as having provided a "strong" deduction of the moral law, that is, one that does not implicitly rely on moral premises. Admittedly, this is perfectly compatible with Kant being confused about the kind of deduction he could offer, which as I mentioned is one of the ways we might read the Henrich interpretation, but exegetical charity dictates that we should adopt an alternative interpretation if one that is at least as textually sensitive can be found.

[50] Moreover, Kant claims in the *Groundwork* that the possibility of hypothetical imperatives "requires no special discussion" (G 4:417), which might suggest that Kant likewise does not regard the claim that we are empirical practical reasoners as requiring a philosophical defense.

Let us examine more closely Kant's worry that the preliminary argument contains a "circle." On my view, the central problem with the preliminary argument, and the basis for Kant's concern, is best characterized in terms of Kant's distinction between negative and positive transcendental freedom. Indeed, as I will show, Kant's own explanation for the circle, as well as his attempt to overcome it, demonstrate a sensitivity to this distinction.

The discussion that introduces the issue of a "circle" consists of five paragraphs, with Kant making the worry explicit in the third paragraph. It begins with a restatement of the conclusion of the preliminary argument:

> [W]e find that…we must assign to every being endowed with reason and will this property of determining himself to action under the idea of his freedom.
>
> But there also flowed from the presupposition of this idea consciousness of a law for acting: that [maxims] must always be so adopted that they can also hold…universally as principles, and so serve for our own giving of universal laws. (4:449)

This is the closest Kant comes to developing the argument in the way I claimed would be most natural. It is stated outright that (i) necessarily, if some being is a rational agent, then that being acts under the idea of freedom. Moreover, in connection with the preliminary argument, it is implied that (ii) necessarily, if an agent acts under the idea of freedom, then that agent is bound by the Formula of Universal Law. Putting these claims together, and keeping in mind Kant's assertion that to act under the idea of freedom is to be "free in a practical respect," yields the following structure of *a priori* entailments:

rational agency → transcendental freedom → moral obligation.

The entailment between rational agency and transcendental freedom is grounded in practical reflection, and is put forth by Kant as a synthetic, *a priori* truth. The entailment between transcendental freedom and moral obligation is put forth as analytic. But in light of Kant's distinction between

negative and positive freedom, how should the above concept of freedom be interpreted? We might interpret the claim in terms of *negative* freedom:

rational agency → negative freedom → moral obligation.

I believe that Kant has negative freedom in mind in the passage that immediately follows:

> But why, then, ought I to subject myself to this principle and do so simply as a rational being, thus also subjecting to it all other beings endowed with reason? I am willing to admit that no interest *impels* me to do so, for that would not give a categorical imperative; but I must still necessarily *take* an interest in it and have insight into how this comes about; for this "ought" is strictly speaking a "will" that holds for every rational being under the condition that reason in him is practical without hindrance; but for beings like us – who are also affected by sensibility, by incentives of a different kind, and in whose case that which reason by itself would do is not always done – that necessity of action is called an "ought," and the subjective necessity is distinguished from the objective.
>
> It seems, then, that in the idea of freedom we have actually only presupposed the moral law, namely the principle of the autonomy of the will itself, and could not prove by itself its reality and objective necessity… (4:449)

In the first paragraph of this passage, Kant introduces the problem of explaining the "interest" we take in morality. In the next paragraph, Kant for the first time makes explicit his concerns about circular reasoning in the preliminary argument. To appreciate the significance of the first paragraph, it is worthwhile to interpret Kant's remarks in terms of our capacity for negative freedom. Kant writes, "I am willing to admit that no interest *impels* me to [obey the moral law], for that would not give a categorical imperative" (ibid.). He then explains that since the categorical imperative expresses an obligation (in his words, an "ought"), obeying the categorical imperative requires free action on our part. But this, in turn, implies that we have the freedom to act contrary to our moral requirements, for "that which reason by itself would do is not always done" (ibid.). In other words, we exercise our negative freedom of choice [*Willkür*] regardless of whether we act from the representation of moral duty. Thus, an agent who regards her capacity for choice as free from external causal determination need not, on that basis, make a practical distinction between the principle of self-love and the Formula

of Universal Law. For she is not causally determined to adopt either as the principle by which she determines herself to action, and so her presupposition of negative freedom does not bear on the question of which principle she should adopt.

Although Kant does not develop this argument beyond the suggestive remarks cited above, it bears mentioning that a capacity for negative freedom is even compatible with a total inability to deliberate and act from the representation of moral duty. For a negatively free agent might still possess an "empirically conditioned" will. As a negatively free agent, she would not be causally determined to act from any particular empirical desire or incentive. Nevertheless, she could not transcend her sensible nature altogether by electing to act on principles of pure reason. Her will would in that respect be "conditioned" by her sensible desires, inasmuch as her "menu" of motivational options would be restricted to those based in sensibility. Kant takes up the possibility of such an agent in the first *Critique*'s "Canon of Pure Reason":

> [I]f the conditions for the exercise of our free choice [*Willkür*] are empirical, then in that case reason can have none but a regulative use, and can only serve to produce the unity of empirical laws, as, e.g., in the doctrine of prudence the unification of all ends that are given to us by our inclinations into the single end of **happiness** and the harmony of the means for attaining that end constitute the entire business of reason…. (KrV A800/B828, my emphasis)

The phrase "conditions for the exercise of our free choice" suggests the possibility of empirically conditioned, negatively free agency. Here the function of practical reason is simply to integrate the various ends to which sensible inclination directs it into a conception of happiness that serves as the will's most general pursuit. This occludes the possibility of a "pure" rational motive, such as the lawgiving form of one's maxim. The logical possibility of such agency entails that the inference from negative freedom to moral obligation is invalid. Moreover, the logical possibility of such agency

likewise entails that there is no analytic entailment between negative and positive freedom.[51] Consequently, the entailment structure

rational agency → negative freedom → positive freedom → moral obligation

is invalid.

Returning to the passage from *Groundwork* III, we see that after raising the topic of our "interest" in morality, Kant first introduces the worry about circular reasoning:

> It seems, then, that in the idea of freedom we have actually only presupposed the moral law, namely the principle of the autonomy of the will itself, and could not prove by itself its reality and objective necessity…. (4:449)

Given the failure of explaining the "interest" we take in morality in terms of the presupposition of our negative freedom, this comment can seem like a non sequitur. After all, that interpretation of the preliminary argument wasn't *circular*; it was *invalid*. But a clue is provided by Kant's claiming that in the preliminary argument, the moral law, as "the principle of the autonomy of the will," was presupposed in "the idea of freedom."

Given the context of the passage, we can think of this claim as expressing, in highly abbreviated form, something like the following thought: The fact that we presuppose our negative freedom in action cannot account for the interest we take in morality. Therefore, when we conclude that we are morally obligated on the basis of the claim that, as rational agents, we act under the idea of freedom, we must be including in the "idea of freedom" not just our negative freedom but also our *autonomy*, or positive freedom. That is, we are implicitly interpreting the preliminary argument as establishing the following *a priori* entailment structure:

rational agency → positive freedom → moral obligation.

[51] See chapter three, footnote 25.

But to say that we presuppose our positive freedom in action is just to say that we recognize the moral law, the principle of our positive freedom, as the principle of our action. Thus, for the purposes of a deduction of the moral law, there is a covert circle in the claim that we act under the idea of freedom.

Admittedly, this is likely to strike the reader as a highly speculative, even strained, interpretation of these remarks. The real basis for it is an explanation of the circle that Kant provides two paragraphs later:

> It must be freely admitted that a kind of circle comes to light here from which, as it seems, there is no way to escape. We take ourselves as free in the order of efficient causes in order to think ourselves under moral laws in the order of ends; and we afterwards think ourselves as subject to these laws because we have ascribed to ourselves freedom of the will: <u>for freedom and the will's own lawgiving are both autonomy and hence reciprocal concepts, and for this reason one cannot be used to explain the other or furnish a ground for it but can at most be used only for the logical purpose of reducing apparently different representations of the same object to one single concept</u> (as different fractions of equal value are reduced to their lowest expression). (4:450, my emphasis)

This passage shows that the basis for Kant's worry is the fact that "freedom and the will's own lawgiving are both autonomy and hence reciprocal concepts," that they are "apparently different representations of the same object," and hence that "one cannot be used to explain the other." According to this line of thought, to say that we necessarily act under the idea of our (positive) freedom just *is* to claim that we necessarily take an interest in the moral law, precisely because the moral law is identical with "the principle of the autonomy of the will itself" (4:449). For that reason, the preliminary argument cannot properly be called a "deduction" of the moral law. A proper deduction of the moral law would have to validate the synthetic, *a priori* proposition that we are bound by the moral law as a categorical imperative, by tracing that claim to some deeper source in our cognition. But that deeper source cannot be the fact that we necessarily act under the idea of our positive freedom, for that is just to claim that we do in fact regard the lawgiving form of our maxims as an unconditional standard of action.

Now, as it stands, this argument is invalid. This is because the term 'under the idea of X' creates an intensional context, i.e., one where substitution of co-referring terms for 'X' is not necessarily truth preserving. (I can be aware that I am drinking water without being aware that I am drinking H_2O.) This suggests the possibility of representing myself in action as an autonomous being without also presupposing the categorical imperative. Perhaps I can *discover*, through Kant's argument, that my conception of myself as an autonomous agent entails a commitment to moral principles as the principles of my autonomy.

I believe that Kant takes up this very possibility when he claims, in his discussion of the circle, that "[w]e do indeed find that we can take an interest in a personal characteristic that brings with it no interest at all in a condition" (G 4:450). Here again, Kant is appealing to practical reflection. However, in this case, Kant specifically appeals to reflection on our capacity for *pure* practical thought and action. He is concerned to show that our conception of ourselves as autonomous has its basis in the self-consciousness of moral agency. What Kant is saying here is that each of us, reflecting on our own agency, is conscious of an ability to be motivated by a consideration that is not grounded in sensible desires: that is what it means to "take an interest" that is not an interest "in a condition," i.e., not an interest in a *sensible* condition. But since, for Kant, the capacity to be so motivated is autonomy, Kant is referring to the self-awareness of autonomous agency, i.e., the awareness that one is able to determine oneself to act on the basis of pure thought. Crucially, Kant claims that this awareness cannot constitute the basis of a deduction of the moral law, since it is grounded in our recognition of the moral law's authority. In particular, Kant describes this self-awareness as the taking of an interest in a "personal characteristic" (4:450). In the subsequent discussion, we find out that this personal characteristic is the "mere worthiness to be happy"—he has in mind here a specifically *moral* worthiness—and hence that "this judgment [that we can take an interest that is not an interest in an

empirical condition] is in fact only the result of the importance we have already supposed belongs to the moral law" (ibid.).

If we compare this passage to a previously cited passage from the second *Critique*, in which Kant asserts that our "cognition" of positive freedom is grounded in our recognition of the moral law as an unconditional practical principle, we see Kant make a nearly identical point:

> But how is consciousness of that moral law possible? We can become aware of pure practical laws just as we are aware of pure theoretical principles, by attending to the necessity with which reason prescribes them to us and to the setting aside of all empirical conditions to which reason directs us. The concept of a pure will arises from the first, as consciousness of a pure understanding arises from the latter. (5:30)

Here the representation of our autonomy is described in terms of the origin of our concept of a "pure will," which arises when we become cognizant of the "setting aside of all empirical conditions" (ibid.). In the *Groundwork* it is described instead in terms of an "interest" we find in ourselves that is not an interest in a sensible condition. Given the evident similarities of these passages, I submit that the respective accounts of how we first become conscious of our autonomy are the same. In both cases it might appear that Kant is setting forth the temporal order of the awareness of our autonomy with respect to our recognition of the authority of the moral law, i.e., that *first* we recognize that we are morally obligated and on that basis *infer* that we are free. But according to the interpretation of moral self-consciousness offered in the previous chapter, we should not read Kant in this way. It is rather the case that there is a single, basic capacity—the activity of pure practical reason, whose internal principle is the Formula of Universal Law—and that the agent's pure apperception of this activity is a constitutive feature of that very activity. If this is right, it implies that the theory of moral consciousness Kant advances with the doctrine of the fact of reason is already present in the *Groundwork*. In other words, already at the time of the *Groundwork*, Kant held that practical deliberation and action includes an original consciousness of the overriding authority of the moral law, and that our most basic awareness of ourselves as autonomous is identical with this moral consciousness. Thus,

the pure apperceptive account of moral self-consciousness advanced in the second *Critique* does not constitute, by itself, any reversal of the position he already espoused in the *Groundwork*.

We should therefore not conclude that Kant's worry about a "circle" in the preliminary argument is based on an illicit substitution of synonymous terms. Rather, Kant's point is that our very notion of autonomy is given content through the pure apperceptive consciousness that attends pure practical thought and action, so we have no conception of what it would be to act under the idea of (positive) freedom outside of moral agency. Therefore, appealing to the claim that we act under the idea of positive freedom cannot supply a non-moral basis for regarding ourselves as morally obligated, and there is indeed "a kind of circle" in the preliminary argument of *Groundwork* III.

Kant tries to overcome this circle with a "deduction of the concept of freedom," which offers a non-moral justification for rational agents to regard themselves as autonomous, or positively free (4:447). In brief, that justification is our status as "intelligence[s]" and therefore, according to Kant, members of an "intelligible" world, one governed by laws of pure reason (4:452). Kant explicates the notion of an "intelligence" by pointing to a distinction in the kind of spontaneity evinced by the understanding and theoretical reason, respectively. While the understanding is drawn into operation by sensibility, and is in that respect sensibly conditioned, theoretical reason produces completely from itself ideas that serve a regulative function in theoretical inquiry. We can think of this distinction as the theoretical analog of the distinction between empirically conditioned, negatively free agency and autonomy. It is because of our capacity for pure theoretical thinking that, according to Kant, we regard ourselves as intelligences, and this is taken to provide a non-moral justification for supposing that we are autonomous:

> As a rational being, and thus as a being belonging to the intelligible world, the human being can never think of the causality of his own will otherwise than under the idea of freedom; for, independence from the determining causes of the world of sense (which reason must always ascribe to itself) is freedom. With the idea of freedom the concept of *autonomy* is now inseparably combined, and with the concept of autonomy the

> universal principle of morality, which in idea is the ground of all actions of rational beings, just as the law of nature is the ground of all appearances. (4:452-453)

Note that in this passage Kant makes a distinction between "the idea of freedom" and "the concept of *autonomy*," claiming that the latter is "now inseparably combined" with the former. Since positive freedom is autonomy, by the "idea of freedom" Kant must here mean only the idea of *negative* freedom, and this is indeed implied by the negative characterization of "freedom" provided by the notion of "independence from the determining causes of the world of sense" (ibid.). The first sentence of this passage recalls the preliminary argument's conclusion that as rational agents we necessarily act under the idea of freedom. Crucially, this "deduction of freedom" *relies on the preliminary argument* to the extent that it is addressed to those who take themselves to be negatively free rational agents. That is, the "deduction of freedom" doesn't eschew the preliminary argument, but rather supplements it, by trying to "bridge the gap," so to speak, between negative and positive freedom. It does so by appealing to the "pure" spontaneity evinced by our capacity for theoretical reason, on the basis of which Kant believes that as rational agents we must regard our wills as not just negatively free but, moreover, autonomous.

Of course, this argument doesn't succeed, as even Kant must have realized by the time he declared in the second *Critique* that a deduction of the moral law is impossible. This is because Kant has given us no reason to suppose that a being capable of pure theoretical reason must likewise be capable of pure practical reason. Kant himself suggests at one point the possibility of a being whose reason does "not break into *practical use*" (4:395). But if such a being is possible, might there likewise exist a being whose theoretical reason is pure but whose practical reason is empirically conditioned? And if so, why couldn't someone who regards herself as a pure "intelligence" *qua* knowing subject regard her will merely as negatively free? Barring a convincing reply to this objection, we must judge Kant's "deduction of freedom"—and, by extension, the greater deduction of the moral law of which it is a component—a failure.

But I do not wish to fixate on this failure. Rather, in conclusion of this chapter, I hope to draw the reader's attention to the specific aim of Kant's argument, the method by which the argument tries to achieve this aim, and what this implies about the "need" to provide a deduction of the moral law.

Groundwork III tries to in some sense *prove* that rational human beings are truly morally obligated—that morality is not just a matter of superstition, an artifact of our sensible nature, or a system of arbitrary customs and conventions. I've argued that part of what motivates this argument is the fact that we can adopt a theoretical standpoint on the world, one from which we can intelligibly ask whether the moral standards that structure our practical deliberation are indeed unconditionally obligating for rational agents like ourselves. By providing a "deduction of morality," we can perhaps allay our concerns that morality is a mere "phantom of the human imagination" (4:407).

But as we saw in the previous section's discussion of the "fact of reason" doctrine, our entitlement to regard ourselves as morally obligated is restricted to the practical use of reason, and the qualification given by the *Groundwork*'s argument from freedom "in a practical respect" entails that, *even if that argument were successful*, the scope of its conclusions would be likewise restricted. In both the *Groundwork* and the second *Critique*, Kant states unequivocally that our entitlement to regard ourselves as moral agents involves no extension of theoretical knowledge beyond appearances. Furthermore, the fact that this entitlement is restricted to reason in its practical use entails that such an entitlement is in any case *compatible* with theoretical doubts we might raise about the reality of moral obligation and our own capacity for moral agency—the very theoretical doubts that motivate the felt need to provide a deduction of morality in the first place.

The above observation is underscored by the fact that throughout Kant's attempted deduction of morality, he makes repeated appeals to items of self-cognition that by his own lights are known only through reflection on our practical capacities. These include the assumption that we are rational agents, as well as the claim that our rational agency involves a presupposition of transcendentally

negative freedom—a conclusion, I've argued, that even his supplementary "deduction of freedom" makes use of. In other words, the argument of *Groundwork* III attempts to legitimize the claim that we are moral agents by appealing to features of our agency that the pure apperceptive structure of such agency enables us to cognize. But as I have just argued, the *Groundwork* already contains within it the pure apperceptive account of moral consciousness that receives fuller articulation with the second *Critique*'s "fact of reason" doctrine.

If that is right, then the deduction Kant attempts in *Groundwork* III is arbitrary in both its aims and methods. It attempts to provide a non-moral basis for regarding oneself as a moral agent, but in doing so it assumes a capacity for non-moral agency. However, while *some* attenuated capacity for agency might be an item of theoretical knowledge—presumably, we can know through empirical apperception that our desires and subsequent behavior tend toward a coherent ordering, in such a way that the latter can be explained by the former—we can nevertheless ask ourselves, when we adopt the theoretical standpoint on the world, whether we in fact possess a capacity for negatively free choice [*Willkür*]. Why, then, does the *Groundwork* contain no attempt to provide a "deduction" of the claim that we are negatively free, resting content with the deliverances of practical reflection? For as we saw, transcendental freedom is treated as a merely potential "property" of the will, and it is a synthetic, *a priori* claim that any rational agent is negatively free. Likewise, at crucial junctures the argument relies on practical reflection to supply a foundation of substantive self-cognition. Yet, despite *Groundwork* III affirming that we are conscious of an ability to "take an interest" in morality, that chapter does not cite our pure apperception of moral agency as a basis for regarding ourselves as autonomous. However, if we are entitled, on the basis of practical reflection, to claim that every rational agent presupposes her transcendentally negative freedom in practical deliberation and action, why aren't we likewise entitled, on the basis of practical reflection, to affirm that in our practical deliberation we are

conscious of the imperatival force of morality, and to make such consciousness the basis for a "critique" of empirical practical reason?

In light of the previous section, I hope to have shown that it is the *Groundwork*'s method, and not the second *Critique*'s, that constitutes a departure from the core methodological commitments of Kantian critique. In aiming to provide a "critique of pure practical reason," the *Groundwork* attempts to meet a demand that has its basis in the theoretical standpoint on the world. Nevertheless, in doing so it does not aim to provide a *theoretical* justification of morality. Rather, it presupposes a capacity for reflective self-cognition and makes *selective* use of our capacity for practical reflection, attempting to vindicate moral agency through appeals to non-moral agency. Moreover, the assumption of a capacity for reflective self-cognition itself has no theoretical justification, but instead issues from the deeper presupposition of the normative autonomy of human reason. We can therefore regard *Groundwork* III as an incomplete expression of the idea that practical reason contains the standard for its own critique. In the *Groundwork*, Kant appeals to the self-consciousness of autonomous agency only to identify a problematic "circle" in a preliminary attempt to find a justification for morality. By contrast, in the *Critique of Practical Reason*, Kant rejects the demand to vindicate pure practical reason on the basis of anything but itself, allowing that the self-consciousness of moral agency constitutes all the "proof" one needs.

CONCLUSION

In this work I have tried to defend Kant's doctrine of the fact of reason against the charge that it constitutes a lapse into pre-critical dogmatism. The basis for my defense was an account of the relationship of pure apperception to the method of critique. Pure apperception is our consciousness of *actively engaging* in a particular rational activity, whether theoretical or practical. In order for a subject to regard the principle of a particular rational activity as a principle of her own reason, her application of that principle in theoretical or practical deliberation must constitutively identify her as the agent of that activity. But this implies that an autonomous rational activity necessarily exhibits pure apperceptive form. The pure apperception of a rational activity includes an at least implicit awareness of its rule or "function," the nature of which must be available for reflective articulation in an effort to achieve explicit self-cognition. As a consequence, the presumption of the autonomy of reason includes the presumption of a capacity for reflective self-cognition. Reason's self-critique, insofar as it presupposes and exhibits the autonomy of reason, is thus a reflective procedure in conception.

The pure apperceptive account of moral consciousness was relevant to the defense of two principal components of Kant's fact of reason doctrine: (i) Kant's claim that a deduction of the moral law is impossible; and (ii) Kant's claim that the fact of reason nevertheless entitles us to regard ourselves as moral agents. With respect to (i), I argued that the self-consciousness of autonomous agency is the pure apperception of specifically *moral* deliberation and action. For this reason, we cannot invoke such consciousness as the ground of a possible deduction of the moral law: our consciousness of ourselves as autonomous agents *just is* our consciousness of adopting maxims on the basis of their suitability for universal adoption. With respect to (ii), I argued that Kant's invocation of the "fact" of moral consciousness indicates the reliance on a specifically *practical* mode of reflection. But the second *Critique*'s reliance on practical reflection is of a piece with a suitably generalized conception of the critical method. Just as the first *Critique* was made possible by reflecting on the conditions of our

theoretical cognition of objects, so the second *Critique* proceeded on the basis of a form of reflection grounded in the pure apperception of moral deliberation and action.

Relative to the concern that the second *Critique*'s fact of reason doctrine is dogmatic, the *Groundwork*'s attempt to offer a deduction of morality will inevitably seem like the superior approach. For regardless of how flawed that argument is, we can nevertheless assure ourselves that it at least *tries* to establish that we are moral agents. As long as we frame our discussion of the fact of reason in these terms, any attempt to vindicate that doctrine will adopt a defensive posture. But as I hope to have shown, it is actually the *Groundwork*'s attempted vindication of morality that constitutes a significant departure from the critical method. The *Groundwork* freely grants that we are conscious of moral requirements in practical deliberation, but it attempts to validate our self-conception as moral agents against doubts that have their basis in a theoretical representation of the world. This, I showed, involves a partial but nevertheless deeply *uncritical* capitulation to a "realist" standard of justification, on the basis of which our capacity for the reflective self-cognition of moral agency is not assumed. The *Critique of Practical Reason*, by contrast, eschews the demand to answer theoretical doubts about our status as moral agents. In its rejection of this demand, it expresses the autonomy of reason.

LIST OF WORKS CITED

Allison, Henry. *Kant's Theory of Freedom*. New York: Cambridge University Press, 1990.

Ameriks, Karl. "Kant's Deduction of Freedom and Morality." *Journal of the History of Philosophy* 19, 1981.

_____. *Kant's Theory of Mind: An Analysis of the Paralogisms of Pure Reason*. Oxford: Clarendon Press, 1982.

Boyle, Matthew. "Kant on Consciousness and the Possibility of Cognition" Unpublished.

_____. "Kant's Categories as Concepts of Reflection." Unpublished.

Bristow, William. *Hegel and the Transformation of Philosophical Critique*. New York: Oxford University Press, 2007.

Carruthers, Peter. *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford: Oxford University Press, 2011.

Dickerson, A.B. *Kant on Representation and Objectivity.* New York: Cambridge University Press, 2004.

Engstrom, Stephen. Introduction. *Critique of Practical Reason*, trans. Werner Pluhar, pp. xv-liv. Indianapolis, Indiana: Hackett Publishing Company, Inc., 2002.

Fix, Jeremy David. "Intellectual Isolation." *Mind* 127(506), 2018.

Franks, Paul. *All or Nothing: Systematicity, Transcendental Arguments, and Skepticism in German Idealism.* Cambridge, MA: Harvard University Press, 2005.

George, Rolf. "Kant's Sensationism." *Synthese* 47(2), 1981.

_____. "*Vorstellung* and *Erkenntnis* in Kant." In *Interpreting Kant*, ed. Gram, pp. 31-39. Iowa City: University of Iowa Press, 1982.

Guyer, Paul. "Kant on Apperception and 'A Priori' Synthesis." *American Philosophical Quarterly* 17(3), 1980.

Henrich, Dieter. "The Concept of Moral Insight and Kant's Doctrine of the Fact of Reason." In *The Unity of Reason: Essays on Kant's Philosophy,* pp. 55-88. Cambridge, Massachusetts: Harvard University Press, 1994.

_____. "The Deduction of the Moral Law: The Reasons for the Obscurity of the Final Section of Kant's *Groundwork of the Metaphysics of Morals.*" In *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, ed. Paul Guyer, pp. 303-341. Lanham, Maryland: Rowman & Littlefield Publishers, Inc., 1998.

_____. "Kant's Notion of a Deduction and the Methodological Background of the First *Critique*." In *Kant's Transcendental Deductions: The Three 'Critiques' and the 'Opus postumum'*, ed. Eckart Förster, pp. 29-46. Stanford: Stanford University Press, 1989.

Herman, Barbara. "On the Value of Acting from the Motive of Duty." In *The Practice of Moral Judgment*, pp. 1-22. Cambridge: Harvard University Press, 1993.

Hume, David. *A Treatise of Human Nature.* New York: Oxford University Press, 2000.

Hurley, S.L. "Kant on Spontaneity and the Myth of the Giving." *Proceedings of the Aristotelian Society* 94, 1994.

Kant, Immanuel. *Anthropology from a Pragmatic Point of View*, trans. Robert B. Louden. In *Anthropology, History, and Education: The Cambridge Edition of the Work of Immanuel Kant*, pp. 227-429. New York: Cambridge University Press, 2007. Abbreviated 'An'.

_____. *Critique of the Power of Judgment*, trans. Paul Guyer and Eric Matthews. New York: Cambridge University Press, 2000. Abbreviated 'KU'.

_____. *Critique of Practical Reason*. In *Practical Philosophy: The Cambridge Edition of the Works of Immanuel Kant*, trans. Mary Gregor, pp. 133-258. New York: Cambridge University Press, 1996. Abbreviated 'KpV'.

_____. *Critique of Pure Reason*, trans. Paul Guyer and Allen W. Wood. New York: Cambridge University Press, 1998. Abbreviated 'KrV'.

_____. *Groundwork of the Metaphysics of Morals*, trans. Mary Gregor. Cambridge: Cambridge University Press, 1997. Abbreviated 'G'.

_____. *The Jäsche Logic*. In *Lectures on Logic: The Cambridge Edition of the Works of Immanuel Kant*, trans. J. Michael Young, pp. 517-640. New York: Cambridge University Press, 1992. Abbreviated 'JL'.

_____. *Lectures on Ethics*, ed. Peter Heath and J.B. Schneewind, trans. Peter Heath. New York: Cambridge University Press, 1997. Abbreviated 'LE'.

_____. *The Metaphysics of Morals.* In *Practical Philosophy: The Cambridge Edition of the Works of Immanuel Kant*, trans. Mary Gregor, pp. 353-603. New York: Cambridge University Press, 1996. Abbreviated 'MM'.

_____. *Notes and Fragments: The Cambridge Edition of the Works of Immanuel Kant*, trans. Bowman, Guyer, and Rauscher. New York: Cambridge University Press, 1997. Abbreviated 'NF'.

_____. *Prolegomena to Any Future Metaphysics*, trans. Gary Hatfield. New York: Cambridge University Press, 1997. Abbreviated 'Prol'.

_____. *Religion Within the Boundaries of Mere Reason*. In *Religion and Rational Theology: The Cambridge Edition of the Works of Immanuel Kant*, trans. Allen Wood and George di Giovanni, pp. 39-215. New York: Cambridge University Press, 1996. Abbreviated 'Rel'.

Kitcher, Patricia. "Kant's Argument for the Categorical Imperative." *Noûs* 38(4), 2004.

_____. *Kant's Transcendental Psychology*. New York: Oxford University Press, 1990.

Kleingeld, Pauline. "Moral Consciousness and the 'Fact of Reason.'" In *Kant's 'Critique of Practical Reason': A Critical Guide*, ed. Andrews Reath and Jens Timmermann, pp. 55-72. New York: Cambridge University Press, 2010.

Korsgaard, Christine M. "Acting for a Reason." In *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, pp. 207-229. New York: Oxford University Press, 2008.

_____. "Morality as Freedom." In *Creating the Kingdom of Ends*, pp. 159-187. New York: Cambridge University Press, 1996.

_____. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press, 2009.

Longuenesse, Béatrice. *I, Me, Mine: Back to Kant, and Back Again*. New York, Oxford University Press, 2017.

_____. *Kant and the Capacity to Judge*. Princeton: Princeton University Press, 1998.

Merritt, Melissa. *Drawing from the Sources of Reason: Reflective Self-Knowledge in Kant's First Critique*. Doctoral Dissertation, University of Pittsburgh, 2004.

_____. "Kant's Argument for the Apperception Principle." *European Journal of Philosophy* 19(1), 2009.

_____. "Varieties of Reflection in Kant's Logic." *British Journal for the History of Philosophy* 23(3), 2015.

O'Neill, Onora. "Reason and Politics in the Kantian Enterprise." In *Constructions of Reason: Explorations of Kant's Practical Philosophy*, pp. 3-27. New York: Cambridge University Press, 1989.

Paton, H.J. *The Categorical Imperative: A Study in Kant's Moral Philosophy*. Philadelphia: University of Pennsylvania Press, 1971.

Pereboom, Derk. "Kant on Intentionality." *Synthese* 77(3), 1988.

_____. "Kant's Transcendental Arguments." *Stanford Encyclopedia of Philosophy*, 2018. https://plato.stanford.edu/entries/kant-transcendental/

Rawls, John. *Lectures on the History of Moral Philosophy*. Cambridge: Harvard University Press, 2000.

Reath, Andrews. "The Categorical Imperative and Kant's Conception of Practical Rationality." In *Agency and Autonomy in Kant's Moral Theory*, pp. 67-91. New York: Oxford University Press, 2006.

_____. "Formal Principles and the Form of Law." In *Kant's 'Critique of Practical Reason': A Critical Guide*, ed. Andrews Reath and Jens Timmermann, pp. 31-54. New York: Cambridge University Press, 2010.

Rödl, Sebastian. "Self-Consciousness and Knowledge." In *Kant Und Die Philosophie in Weltbürgerlicher Absicht: Akten des Xi. Kant-Kongresses 2010*, eds. Ruffing, La Rocca, Ferrarin, and Bacin, pp. 357-370. Berlin: De Gruyter, 2013.

Russell, Bertrand. *Introduction to Mathematical Philosophy*. London: Routledge, 1919.

Scanlon, T.M. *Being Realistic about Reasons*. New York: Oxford University Press, 2014.

Smit, Houston. "The Role of Reflection in Kant's *Critique of Pure Reason*." *Pacific Philosophical Quarterly*. 80(2), 1999.

Strawson, Peter F. *The Bounds of Sense: An Essay on Kant's Critique of Pure Reason*. London: Methuen, 1966.

Sussman, David. "From Deduction to Deed: Kant's Grounding of the Moral Law." *Kantian Review* 13(1), 2008.

Timmermann, Jens. "Reversal or Retreat? Kant's Deductions of Freedom and Morality." In *Kant's 'Critique of Practical Reason': A Critical Guide*, ed. Andrews Reath and Jens Timmermann, pp. 73-89. New York: Cambridge University Press, 2010.

Tolley, Clinton. "Kant on the Content of Cognition." *European Journal of Philosophy* 22(2), 2014.

Willaschek, Marcus. "Die Tat Der Vernunft: Zur Bedeutung der Kantischen These vom 'Faktum der Vernunft'." In *Akten des Siebenten Internationalen Kant-Kongresses*, ed. Gerhard Funke, pp. 455-466. Bonn: Bouvier, 1991.