ONE AND THE SAME REASON

By

Kenneth Edward Dale Walden
A.B. Harvard

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment the requirements for the degree of

Doctor of Philosophy in Philosophy
at the
Massachusetts Institute of Technology
[September 2011]

Author ....
Kenneth Edward Dale Walden
June 8, 2011

Certified by .....
Richard Holton
Professor of Philosophy
June 8, 2011

Certified by ....
Rae Langton
Professor of Philosophy
June 8, 2011

Accepted by..... .....
Alex Byrne
Professor of Philosophy
Chair of the Committee on Graduate Studies

One and the Same Reason

by

Kenneth Edward Dale Walden

Submitted to the Department of Linguistics & Philosophy on 13 June 2011 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Philosophy at the Massachusetts Institute of Technology.

My dissertation is about the relationship between theoretical and practical reason. I argue that these two kinds of reason are unified in important respects. In Chapter One I argue that there is a single, fundamental kind of reasoning (roughly, unrestrained self-reflection) and that theoretical and practical reason ought to be understood as instances of this more fundamental kind of reasoning, distinguished only by their subject matter. I then argue that two formulations of Kant's Categorical Imperative jointly codify the activity of this basic reasoning. Therefore, the Categorical Imperative is, in this sense, the supreme principle of reason. In Chapter Two I show how the very abstract norms formulated in Chapter One can be sharpened if we connect them to the conditions of human agency. I argue that the demands of being an agent require us to submit to a procedure of negotiation and legislation with other agents that is similar to the contractualism of Hobbes and Rawls. The difference between my view and theirs is that my contractualism, because it is tied to our agency, issues in categorical requirements. In Chapter Three I develop a theory of normative concepts that satisfies two demands that have appeared incompatible: the demand that our normative concepts be intimately connected to human nature and the demand that normative items be things we aspire to, and thus things that are relevantly *beyond* us and our activities. I show how we can satisfy both of these desiderata through an open-ended constructivism that understands normative items as transcendent ideals. In Chapter Four I argue that a robust, philosophically serviceable distinction between theoretical judgments about the world and practical judgments about what one ought to do cannot be sustained because these two kinds of judgments are inextricably entangled. They are entangled because we must employ both kinds of judgment to fully explain actions. This fact entails that practical and theoretical judgments occupy a single holistic theory.

Thesis supervisor: Richard Holton
Title: Professor of Philosophy

Thesis supervisor: Rae Langton
Title: Professor of Philosophy

If pure reason of itself can be and really is practical, as the consciousness of the moral law proves it to be, it is still only one and the same reason which, whether from a theoretical or a practical perspective, judges according to *a priori* principles; it is then clear that, even if from the first perspective its capacity does not extend to establishing certain propositions affirmatively, although they do not contradict it, as soon as these same propositions belong inseparably to the practical interest of pure reason it must accept them. — Kant, *Critique of Practical Reason*

Carnap has maintained that this is a question not of matters of fact but of choosing a convenient language form, a convenient conceptual scheme or framework for science. With this I agree, but only on the proviso that the same be conceded regarding scientific hypotheses generally. Carnap has recognized that he is able to preserve a double standard for ontological questions and scientific hypotheses only by assuming an absolute distinction between the analytic and the synthetic; and I need not say again that this is a distinction which I reject. The issue over there being classes seems more a question of convenient conceptual scheme; the issue over there being centaurs, or brick houses on Elm Street, seems more a question of fact. But I have been urging that this difference is only one of degree, and that it turns upon our vaguely pragmatic inclination to adjust one strand of the fabric of science rather than another in accommodating some particular recalcitrant experience. — Quine, "Two Dogmas of Empiricism"

The world is nothing, the man is all; in yourself is the law of all nature, and you know not yet how a globule of sap ascends; in yourself slumbers the whole of Reason; it is for you to know all, it is for you to dare all. — Emerson, "The American Scholar"

The universe is God itself, the universal outpouring of its soul. — Chrysippus

By the way, would you convey my compliments to the purist who reads your proofs and tell him or her that I write in a sort of broken-down patois which is something like the way a Swiss-waiter talks, and that when I split an infinitive, God damn it, I split it so it will remain split, and when I interrupt the velvety smoothness of my more or less literate syntax with a few sudden words of barroom vernacular, this is done with the eyes wide open and the mind relaxed and attentive. The method may not be perfect, but it is all I have. — Chandler

Contents

Front matter

Introduction

My dissertation is nominally about the relationship between theoretical and practical reason. And while I do argue for two theses that can both be understood as affirmations of the 'unity of reason', I also stray from this narrow path. The point about unity is important because it offers an antidote to a mistake in work on the foundations of ethics and inquiry. Participants in these debates have labored under the illusion that we have two masters: The World demands one kind of fealty and The Good demands another. A lot of philosophy—done by realists, naturalists, expressivists, and constructivists alike—has aimed to show how we can serve both masters at once. My goal was to expose the speciousness of the dichotomy underwriting these projects.

I quickly learned that philosophers are more easily liberated from their ideological chains in theory than in practice. So instead of proselytizing, I turned to a more solitary enterprise. I set about building a system. The two "unity" theses I champion emerge as corollaries in this system.

Our ability to doubt, demur, reflect, and scrutinize is the core activity of reason. That we can, in principle anyway, do this in an unbounded way is what distinguishes human rationality from the meager rationality of animals. This ability is also the source of both human freedom and normative thought. It is because we can scrutinize our inclinations, prejudices, social situation, and other contingent forces on our behavior so thoroughly that we can liberate ourselves from them and be free. And it is for the same reason that we can legitimately ask not just *what shall I do, given my inclinations prejudices, social situation, and other contingent forces?* but *what ought I do?* I may feel the force of an inclination, but the question of whether that inclination gives me

a reason to act is one opened when detach myself from it.

I am not the first person to argue for a close connection between reflection, freedom, and normativity. Christine Korsgaard's version of this thesis in her book *The Sources of Normativity* is the most recent, and probably most familiar. But what Korsgaard neglects is how important it is that reason's power of reflection be unrestricted, unlimited, and open-ended. This point is the essence of my way of thinking about normativity. If reason's power to challenge our attitudes is circumscribed, then scrutinizing those attitudes with reason becomes a mere matter of checking them against a finite set of higher-order attitudes. This is problematic for several reasons. It means that once we achieve equilibrium amongst this circumscribed set of attitudes, there is no further question about whether the conceptions of the Right and the Good entailed by these attitudes are correct. It identifies reason, which was supposed to be the source of freedom, with a passive state of equilibrium. It enables a kind of unfortunate subjectivism: different people may satisfy the circumscribed demands of reason in different ways, and this means that different notions of the Right and the Good may hold for them.

These problems can be overcome by understanding the scrutiny applied by reason as open-ended. In chapter three I explain what I mean by open-ended. I understand the claim that the demands of reason are open-ended as following from some surprising set-theoretic properties of perspectives. Reason demands that we subject our judgments to the scrutiny of ever more perspectives (in the sense that if a judgment would not be held from that perspective, that gives us a *prima facie* reason to refrain from making that judgment). There is no totality of all perspectives because the concept *perspective* is indefinitely extensible in the way that Michael Dummett says the concept *set* is. From this follows the open-endedness of reason.

What results from this account of the activity of reason, when paired with the approach to normativity outlined above, is a conception of normativity that embraces neither horn of Socrates's famous *Euthyphro* dilemma. On this view what is Right and Good is neither something independent of us and our activities, nor something determined by us and those activities.

In chapter one I connect the open-endedness of reason to questions about the

objectivity of both science and ethics. The dialectic of this chapter is a little unfortunate, since it was written when my main concern was demonstrating the unity of reason. The basic argument is something like this. The demand for objectivity in science arises not from facts about the world (as Bernard Williams has argued), nor from any kind of manifest intuition (as sundry lazy people have argued). It comes from reason's demand that we subject scientific judgments to the scrutiny of ever more perspectives. But reason has precisely the same power to scrutinize *ethical* judgments. So the very same objectivity is demanded in both science and ethics, and for the very same reason.

So construed, the activity of reason produces two requirements, each of which we can think of as codifying reason's activity. These principles have authority over us insofar as they are constitutive of reasoning, and it is our ability to guide our actions through reasoning that gives normative talk its sense in the first place. The first tells us that disagreement between perspectives on a judgment (practical or theoretical) is a *prima facie* reason to revise that judgment. The second says that total concordance of judgments is a regulative ideal for our projects. When restricted to just practical judgments, these two requirements come very close to the second and third formulations of Kant's Categorical Imperative, the Formula of Humanity and the Formula of the Realm of Ends. Thus, what I show in this chapter is that a generalization of the Categorical Imperative is the most fundamental principle—the constitutive principle—of *all reason*.

I said that one of the requirements of reason refers to a regulative ideal of total and systematic concordance. Chapter three includes a discussion of the nature and function of this ideal. Importantly, I say that this is a genuine ideal, in Kant's sense of the term. It is a state that is *in fact* and *in principle* unachievable. Nonetheless, it can guide our projects in the following way. We can think of ethical theory and epistemology as engaged in building models of their target normative terminology: right, just, good, justified, knowledge, and so on. These models represent a compromise between two things. The first is the requirement that our normative terminology be implementable: that we can have a conception of the right, justice, the good, justification, and knowledge that is definite enough that it can regulate our inquiries in the

usual way ("the system of apartheid is unjust"; "don't make that inference, it would be unjustified"). The second is the demand that these conceptions hew as closely as possible to our unattainable ideal of them. I suggest that we find this kind of methodology in the natural sciences. Scientists build models that are a compromise between the pure virtues of explanation, things like order, symmetry, uniformity, (virtues which we think are not wholly achievable in the actual world) and the messiness of their actual data.[1]

The requirements I sketch in chapter one are very abstract. In chapter two I try to show how more specific normative requirements arise. The strategy in chapter one was to connect normative requirements to the constitutive standards of reasoning. The idea here is to connect such requirements to the constitutive standards of *agency*. On my view, agency and reasoning are related in the following way: agency is the particular way that human beings (and their particular social and physical dimensions) manifest the ability to regulate their behavior through reasoning. Thus the demands of agency are still categorical for actual human beings, though this categoricity is not as modally robust as that of the norms connected to reasoning, which are categorical for all rational creatures, whatever their nature. We should think of the norms constitutive of agency as more precise *implementations* of the norms attached to reason, implementations that are the necessary way for creatures like us to uphold the laws of reason, though not necessarily for other rational creatures. I therefore hold a two-tiered version of a view usually called constitutivism (one propounded in the work of David Velleman and, again, Christine Korsgaard): very general norms arise from the conditions of reasoning, and more specific versions of those norms arise from the particular way that human beings in particular must carry out that reasoning.

The particular view of agency I offer goes like this. The nature of agency is something that human beings create through their actions. What agency is, is something

[1]Aside from his completely inappropriate substitution of "system" for "systematicity", Quine (1981) puts the point best: "A good scientific theory is under tension from two opposing forces: the drive for evidence and the drive for system. Theoretical terms should be subject to observable criteria, the more the better, and the more directly the better, other things being equal; and they should lend themselves to systematic laws, the simpler the better, other things being equal. If either of these drives were unchecked by the other, it would issue in something unworthy of the name of scientific theory: in the one case a mere record of observations, and in the other a myth without foundation."

constituted by the actions of you, me, and everyone else. A consequence of this arrangement is that people are committed to a kind of contractualist scenario of cooperation and negotiation in virtue of the demands of agency. They are so committed not because it is in their best interest (as Hobbes says), nor because it enshrines what is reasonable (as Rawls says), nor because it follows from a dogmatic theory of reasons (as Scanlon says) but because our submission to this process is required by agency itself. The more specific norms this constitutivism gives us—including moral norms—are produced by this procedure. My constitutivism's reliance on this kind of procedure gives it a signal advantage over its rivals: it shows how quite specific norms about morality can arise from the superficially thin demands of agency.

The final chapter develops the germ of an idea found in chapter three: that norms are intimately connected to agency and action, that agency is a natural phenomenon, and that this fact scrambles the distinction between theoretical and practical judgments. I argue that we cannot make any clear, philosophically useful distinction between practical and theoretical judgments, between judgments about *what to do* and judgments about *how things are*. This is importantly different from the thesis advanced in chapter one. There I argued that judgments about what to do in the realm of science—notably about what one ought to believe—and judgments about what to do in other walks of life are unified by their having a common source. In brief, I argue that theoretical and practical normativity are unified. Here I argue against the dichotomy that I think underwrites the division between judgments about what to do writ large (normativity) and judgments about how things are (description).

So this chapter deals with a dichotomy more fundamental than the one addressed in chapter one. The argument is modeled on a Quinean indispensability argument: first we show that both theoretical and practical judgments are necessary for some explanatory task, and then we show that this implies that they occupy a holistic 'web'. This latter fact, I argue (again on Quinean grounds) entails that we cannot maintain a philosophically useful dichotomy between judgments about what to do and judgments about how things are.

One upshot of this view is a kind of quietism about normativity. Normativity is immanent in the natural world. It is so not because goodness is identical to some

particular natural property (as 'Cornell realists' maintain), but because we cannot untangle facts and values in a way that allows us to ask meaningfully ask the question, "how are values introduced into a world of facts?" And this, I maintain, leaves very little room for traditional metaethics.

Acknowledgments

I had the great fortune of having a committee of exceptional acuity, generosity, and patience: Richard Holton, Rae Langton, Kate Elgin, and Sally Haslanger. I am grateful to them. Thanks also to Steve Yablo, Julia Markovits, Judy Thomson, Bob Stalnaker, Agustín Rayo, Vann McGee, Alejandro Pérez Carballo, Paolo Santorio, Katia Vavova, Sophie Horowitz, Paulina Sliwa, Tom Dougherty, Mahrad Almotahari, and audiences at fora where some of this work has been presented. Most of all, thanks to Alice Phillips Walden.

Chapter 1

The unity of reason

In this paper I address an old question. What is the relationship between the methods of science and the methods of ethics? Now, I think of science and ethics very broadly. I think of scientific inquiry as Quine did: as encompassing our best ways of understanding how things are, and so including all the methods, axioms, and protocols necessary to that understanding. Science, as an activity, is therefore something very close to what we call 'theoretical reason'. Analogously, I think of ethics as the ancients did, as something very close to practical reason: the study of how we should live, what we should do, and what things matter. So I am interested in two closely related questions: What is the relationship between our methods of finding out how the world is and our methods of figuring out what to do, and how do the attendant forms of normativity for each of these methods (if they are normative) relate to each other?

Conventional wisdom offers two approaches to these questions. One line of thought is that the methods of ethics and science have little in common. There may be some points of overlap, but these will be very abstract and without much content. And when it comes to the nitty-gritty of constructing theories and making plans, the two will be turn out to be, in Stephen Jay Gould's rococo phrase, 'non-overlapping magisteria'.[1] The second approach aspires to a reduction of one thing to the other.

[1]See Gould (2002). Bernard Williams's (1985, ch. 7) view is similar to this, albeit more subtle. Non-cognitivists and those distinguishing between 'theoretical' and 'practical' standpoints may also fall into this camp.

We can say that ethics is no more than a branch of a particular science, like 'human ecology'. Or, conversely, we can say that scientific methodology is no more than a particularly robust means to some practical ends—to the satisfaction of our desires, say.[2]

I want to defend the rudiments of a third option. My view is a version of Kant's. In the *Groundwork*, Kant writes that "there can only be one and the same reason, distinguished solely in its application." Kant thinks that there are not two distinct faculties, theoretical and practical reason, but a single faculty of reason. Some interpreters of Kant even see him holding that this single faculty of reason has a single 'supreme principle' that unifies theoretical and practical reason: the Categorical Imperative.[3]

I argue for a version of this thesis. Neither science nor ethics is a branch of the other, but their methods share a common foundation, and their characteristic forms of normativity have a common source in this foundation. This foundation is our power of critical reason, that is, our ability to detach ourselves from our instincts, inclinations, and other attitudes and hold them up to reflective scrutiny. This power, I argue, is the reason that we have normative thought in the first place. It is because we have this power to detach ourselves from our attitudes that we can ask not just descriptive questions like "what will I come to believe?" and "what am I inclined to do?", but normative questions like "what *should* I believe?" and "what is the *best* thing to do?". This critical power of reason engenders a substantive norm, one that directs the activity of both theoretical and practical reason. This principle is a first cousin of the Categorical Imperative. It requires us to strive to make all judgments that fall under the scrutiny of reason objective, and to regard the convergence of opinion on those judgments as a regulative ideal. In sum, then, I argue that reason is unified in three closely connected ways: our epistemic and practical methods bottom out in the activity of a single faculty of critical reason; this faculty generates both epistemic and

[2]For the idea that ethical normativity should be understood as a small part of the natural world see Owen Flanagan et al. (2007), Richard Boyd (2003), and, amongst scientists, E. O. Wilson (1998, ch. 11). For the latter reduction, of epistemic normativity to practical normativity, see Hilary Kornblith (2002, ch. 5).

[3]See the *Groundwork*, 4:391; cf. the *Critique of Practical Reason*, 5:91. Onora O'Neill (1989, ch. 1) defends the thesis that the Categorical Imperative is the supreme principle of all reason.

practical normativity; and this activity engenders a substantive norm that serves as a common methodological foundation for both science and ethics.

I make a kind of transcendental argument. It begins with a principle that is plainly very important to the conduct of science, one directing us to strive for objectivity in our scientific pursuits. I ask why we are justified in following this principle, and argue that the only way to adequately justify this principle is to show that it arises from the critical power of reason. In particular, a familiar argument for why we are required to seek this objectivity—offered by Bernard Williams as part of an argument for the disunity of reason—does not work. I argue that the only way to justify this principle about objectivity is to accept that there is a critical power of reason as I describe it. But from this follows a corollary: the unity of reason. In short, then, my strategy is to show that we are justified in our present understanding of scientific inquiry only if we accept an account of science which entails my thesis about unity.

1.1 The principle to be justified: the Objectivity Principle

I begin by donning my skeptic's hat and challenging one of science's essential precepts. What I am interested in is described well by Max Planck in a statement about the ambitions of physics.

> The goal is nothing less than the unity and completeness of the system of theoretical physics, and not only with respect to all particulars of the system, but also with respect to physicists of all places, all times, all peoples, all cultures. Yes, the system of theoretical physics demands validity not merely for the inhabitants of this earth, but for the inhabitants of the heavenly bodies.[4]

Planck talks about 'unity' and 'completeness', but I think the idea pressed here is better understood as one about *objectivity*. There are other notions of objectivity of course: being true to the nature of things in themselves, taking up a view from

[4]From *Acht Vorlesungen über theoretische Physik: gehalten an der Columbia University im Frühjahr 1909*, qtd. Daston and Galison (2007).

nowhere, seeing things as God would see them. These notions may or may not overlap with Planck's. But Planck's ambitions, lofty as they are, strike me as something like a minimal standard of objectivity. We may not be able see how the world is in itself, and we may never have a God's-eye view, but we should try to make our picture of the universe as broad-minded and unparochial as possible. What this means is that how things seem to someone else—someone standing on a far-flung planet, someone living in a different time, someone entertaining different assumptions, someone reared in a different cultural medium—should *matter* to us. It should affect which judgments we accept. This is not to say that these other perspectives have overwhelming force, just that they have *some* force in the game of rational deliberation.

I will call this thought, that we really are required to pursue the goal that Planck describes, the *Objectivity Principle*. The principle has two different facets worth flagging. First, it is obvious that divergence of opinion within one person is an error in need of rectification. There is something plainly wrong with Seneca believing that Nero fiddles and that Nero doesn't fiddle. The Objectivity Principle adds to this that divergence across *different perspectives* also amounts to an error:

> **Objectivity Principle, First Facet.** A divergence of opinion between individuals occupying two perspectives, like a disagreement between terrestrial and Martian physicists, represents a kind of error and therefore presents us with a *prima facie* reason to correct that error, for us to restore concordance.

Of course, it may turn out that the physicist from Mars suffers from misleading evidence or cognitive malfunction, and so I ought to hold fast to my opinion and work to convince him. Or it may be that I am the one suffering from this malady and should revise my opinion. Both courses of action are consistent with the thought that divergence is an error, and that, *ceteris paribus*, something needs to be done to fix it.[5] The second facet of the Objectivity Principle is more oriented toward the big picture:

[5]There is a large literature about divergence of opinion and objectivity. The notion of objectivity I broach here is not unlike the procedural conception discussed by Thomas Nagel (1989, pp. 3-4) and Gideon Rosen (1994). It is also in the same spirit as the claim that science exhibits what Crispin Wright (1992b, p. 93) calls 'Cognitive Command'.

> **Objectivity Principle, Second Facet.** A total concordance of opinion, wherein everyone occupying sundry perspectives agrees on a single picture of how things are, is a regulative ideal for science.

This facet of the Objectivity Principle does not give us advice on individual cases. Instead it urges us to work toward this ideal of total concordance in the long run.[6]

My argument for the unity of reason beings by challenging this Objectivity Principle. I argue that various attempts to justify the Objectivity Prinicple in the face of scrutiny fail, and that the only way to overcome this scrutiny is to accept a justification rooted in a single power of critical reason that is both theoretical and practical.[7]

1.2 First attempt: Williams's common world

We know—never mind how—that we are in a room with uniform temperature and one hundred thermometers. Two-thirds of the thermometers read fifty degrees while one-third read sixty. Insofar as we are interested in the temperature of the room, we

[6]I call these different facets of a single principle rather than two principles because I want to leave open the possibility that the second entails the first.

[7]The idea of searching for a justification of the Objectivity Principle may strike some as queer because the principle looks so innocuous. But this assumption is naïve. The Objectivity Principle may *look* like an obvious principle that has always been with us, but there is reason to think, at least according to some historians, that it is actually an invention of the latter part of the Enlightenment—in debates over moral philosophy no less!—and so something about which we can quite coherently ask: is this principle worth adopting? On this subject see Daston (1992) and Daston and Galison (2007). I disagree with some of Daston and Galison's analysis, as I think claims about the novelty of certain conceptions of objectivity fail to appreciate the sophistication of later Greek epistemology. The Pyrrhonists' "method of opposing appearances" in particular presupposes a forerunner of the Objectivity Principle. (See Sextus Empiricus, *Outlines of Pyrrhonism*, I.13-9.) But that's an altogether different issue. Similarly, one might protest that the first facet of the Objectivity Prinicple is tautologous because disagreement between two individuals means that one of them must believe falsely, which means that one of them must have made an error. I think this thought jumps too quickly to a propositional analysis. It seems the sort of reason that we would think that a disagreement between two scientists amounts to their believing incompatible propositions is that we accept something like the Objectivity Principle for scientific discourse. If two people disagree on matters of taste, we do not automatically assume that there are two incompatible propositions at stake, and this is because we are unsure whether the Objectivity Principle holds in the domain of taste. Thus whether the Objectivity Prinicple holds for a domain of discourse is something we have to reckon with as part of the project of assigning content to that discourse. Wright, op. cit. is enlightening on this point.

seem obligated to rectify this situation. We should do something to restore concordance, like recalibrating the thermometers or, at least, investigating whence the discrepancy comes. Why does this requirement exist? It should be obvious enough: thermometers aim to reliably indicate temperature, and we antecedently know the temperature in this room is uniform, so any divergence of opinion amongst the thermometers must represent an error.

The most obvious, most tempting, and most natural way to justify the Objectivity Principle is to scale up this story about thermometers to one about the world as a whole. Instead of a uniform room temperature, we have a single, constant world containing many perspectives, some of which are occupied by human beings. Instead of different locations within this room, we have those perspectives. And instead of thermometers we have the judgments that inquirers make while occupying those perspectives. The Objectivity Principle is then justified in the same way as our requirement to recalibrate the thermometers: there is a single world common to all inquirers, and insofar as our epistemic methods ought to be hemmed in by the shape of the world, we should accept norms reflecting this particular fact about the world.

I will call this program for justifying the Objectivity Principle the *Thermometer Story*. Most ways of justifying the Objectivity Principle that leap to mind will rely on the Thermometer Story in one form or another. If, for instance, we are reliabilists, we might tell a story on which adhering to the Objectivity Principle produces true beliefs. But how would such a story go? Such a story will invariably depend on the world being the way the Thermometer Story says it is. These strategies for justifying the Objectivity Principle in the same way we justify our 'ordinary' beliefs all eventually come back to something like the Thermometer Story.

Whether the Thermometer Story works is closely connected to our question about the unity or disunity of reason. Bernard Williams, for example, seems to have something like the Thermometer Story in mind when he sets out to prise apart science and ethics. "The distinction between science and ethics," he writes, is that "in a scientific inquiry there should ideally be convergence on an answer, where the best explanation of the convergence involves the idea that the answer represents how things are; in the area of ethics, at least at a high level of generality, there is no such coherent hope."

Now, Williams concedes that "it might well turn out that there is convergence in ethical outlook." But that is not a problem for his distinction: "the point of the contrast is that even if this happens, it will not be correct to think that it has come about because convergence was guided by how things actually are, whereas convergence in the sciences might be explained in that way."[8] Williams is talking about explanation rather than justification, of course, but here the two go hand in hand. The crux of Williams's argument for the separateness of ethics and science is that science's obligation to the Objectivity Principle is grounded in its concern with a single world that is the common object of all perspectives. But ethics has no such object, and so there is no reason to expect it to be governed by the substantive norms the world imposes on science. Hence the disunity of science and ethics.

But Williams's conclusion is too hasty, for the Thermometer Story (and with it all 'ordinary' justifications) faces a deep problem. The kind of story we get from ruminating on thermometer recalibration is usually the most effective kind of justification we have. But for our most fundamental and theoretically pervasive methods this approach runs aground on the shoals of circularity. Suppose we want to justify our use of induction. If we wanted to tell a Thermometer Story to do this, we would point to a certain facet of the world—its observational regularity or something like that—and argue that induction is a good method because of this: because the world is observationally regular, induction is a good method. But this is no good at all. Induction is so omnipresent in every bit of our reckoning of the world that we can never be sure that things don't appear observationally regular *just because* we use induction. And this worry persists even if, as a matter of fact, the world is observationally regular. For as long as our observation that the world is observationally regular is essentially conditioned on our use of induction to understand the world, it can give us no independent justification of induction. As long as we cannot shake the worry that we are justifying each method with its own shadow, our justification is incomplete.

The same problem afflicts the Thermometer Story. The Objectivity Principle is partly constitutive of the very epistemic methods we would need to gather evidence for and against the propositions that would make up a Thermometer Story. To see

[8]Williams (1985, p. 136).

this, let's imagine how we would arrive at the one-world proposition at the heart of the latter. I see a beagle before me. Now suppose I become worried about whether this is a real beagle out in the real world, one which other people occupying other perspectives might also espy, or just a bit of beagle imagery haunting me. But now someone else approaches the beagle and announces, "Lo, a beagle!" If I am a normal person, then I will take this as evidence that there is a real beagle, one common to me and this other observer, and that this other observer occupies his own perspective on the beagle that discloses things to him not revealed to me. But the way in which I must be 'normal' to make this inference is vital. In order to understand this observer's testimony in this way, I must already understand him as a fellow inquirer occupying a different perspective on a common world. And this means I must view myself as one point with one perspective amongst many other points, each of these perspectives offering, by its very nature, an *incomplete* view.

Contrast this with someone who is not in this frame of mind. My crazy uncle Shmuel does not believe that there is a single world common to multiple perspectives which different inquirers might occupy. He identifies the world with all that is present before his mind. He does not believe that he has a *perspective* on the world because that would mean that there is a way the world is beyond how he sees it. He may not even countenance the 'perspectives' of different temporal slices of himself. He is a radical solipsist, a solipsist *not just* about other people or other minds, but about a world admitting of multiple incomplete perspectives. Consequently, if Shmuel had our experiences, if he saw beagle imagery and then person-yelling-"Lo, a beagle!" imagery, he would not infer from this that there is another person with a different perspective on the beagle. He might see correlations between the blips and bloops before his mind (depending on whether he is a solipsist about his past and future selves), but none of these correlations will add up to the thought that there is another way to see the world beyond his own. To *us* this seems like best explanation for what we have witnessed, but it is not even in the range of hypotheses Shmuel entertains. He is too far gone to understand it as *evidence for* anything like the propositions of the Thermometer Story.

How do we get Shmuel in the right frame of mind to see this sort of episode as

evidence for a beagle common to multiple perspectives? Well, we must endow him with epistemic methods that will make him afford the testimony of other people the sort of status that we do. He must adopt principles that at once make him recognize certain events as testimony and care about that testimony in a special way. But this just means inculcating Shmuel with the Objectivity Principle: the Objectivity Principle just is the norm that requires people to care about the testimony in the way Shmuel must if he is to understand his evidence in this particular. Thus the Objectivity Principle is, in a certain sense, a precondition on Shmuel's coming to accept anything like the Thermometer Story.[9]

The point I am making with Shmuel's outré example is loosely connected to one from the logical empiricists. The idea, very roughly, is that we must orient ourselves toward the world in a particular way—through the adoption of certain methods, concepts, and schemes—for given modes of inquiry to even be possible. We can revise this orientation, of course but not through the normal course of inquiry: it takes something like a scientific revolution. I am urging a similar story. In order to see certain episodes as evidence for the propositions in the Thermometer Story we must take up a certain orientation. If we lack this orientation, then we need to undergo a revolution of mind that endows us with the appropriate methods. But this revolution will invariably include the Objectivity Principle itself.[10]

This means that the Thermometer Story fails to justify the Objectivity Principle in just the way its imaginary analogue would fail to justify induction. There is an essential circularity in these stories: we cannot come by them without first taking up the principles to be justified. And so we are burdened with a worry about circularity, that the world appears congenial to a justification of our basic methods *just because* we

[9]I don't want to rule out the possibility that the reverse direction also holds. It may be that we cannot deduce the Objectivity Principle without already assuming that there is one world encompassing multiple perspectives. This would mean that the two things come as a package, but that wouldn't disrupt my overall argument.

[10]cf. The relationship between this conception of confirmation and Duhemian holism that I discuss elsewhere. See Michael Friedman (1991) for a careful development of these themes. For the suggestion that growth of science in the Sixteenth and Seventeenth Centuries was attended by a dramatic shift in scientists' conception of the world (apropos my earlier discussion of Daston and Galison), see Alexandre Koyré (1957).

employ those basic methods. This is enough to sink our justification.

The Objectivity Principle cannot be justified by what, at first blush, was the obvious approach: the suggestion that it is an apt principle for scientific inquiry because it codifies an important feature of the outside world. And with the failure of this suggestion comes the likely failure of all the ordinary methods of justification—reliabilism, constructivism, etc.—that I suggested must in some way rely upon it. Two other conclusions follow along too. First, Williams's way of driving a wedge between science and ethics cannot work. Science is not made objective from without, and the Objectivity Principle cannot be justified as if it were. This opens the door to the possibility that ethics may be objective (in some sense) too, and that its objectivity may lie in the same place as science's. And second, if we are to justify the Objectivity Principle, we must try something a little more radical.

1.3 Second attempt: Some transcendental arguments

The special way the Thermometer Story failed might lead us to an alternative way of justifying the Objectivity Principle. The Objectivity Principle was too fundamental and too pervasive to be justified by pointing to a fact about the world. But maybe those very features can be spun into a different kind of justification. Suppose the Objectivity Principle is so fundamental and so pervasive that we cannot find any non-circular positive justification for it, but we just as clearly cannot go on without it. Those two facts might constitute a kind of justification of the Objectivity Principle, one premised not on a feature of the world, but on the Objectivity Principle's indispensable role in our epistemic endeavors.

I call this strategy a transcendental argument. That term is usually associated with the Copernican tactics of the *Critique of Pure Reason*, but if we understand it with sufficient liberality, we see examples sprinkled throughout the history of philosophy. Let me mention just three representatives. First, Aristotle says that the principle of non-contradiction is not susceptible to derivation from prior principles, which means

that an ordinary justification isn't in the cards. But something else is. We can show that the principle of non-contradiction is a first principle, one that is "necessary for anyone to have who knows any of the things that are" because the rejection of non-contradiction always results in a kind of cognitive collapse. And this fundamentality justifies our use of the principle. Second, Hans Reichenbach compares the predicament of justifying induction to the travails of a blind man lost in the mountains. The man comes upon a path he can follow with his walking stick. He has no evidence that the trail will lead him to safety rather than to a den of ravenous bears. But it seems that he is nonetheless justified in following the path because it is the only thing he can do. The same, Reichenbach says, applies to our use of induction: even if we cannot give an ordinary non-circular justification for these methods, we can show that they are *indispensable* to some rationally required enterprise. Third, Wittgenstein muses, "whenever we test anything, we are already presupposing something that is not tested. Now am I to say that the experiment which perhaps I make in order to test the truth of a proposition presupposes the truth of the proposition that the apparatus I believe I see is really there (and the like)?" What Wittgenstein means is that particular inquiries must take certain things for granted, and so even if I do not have positive justification for a proposition or norm, I may be entitled to assume it insofar as I am entitled to undertake that inquiry. These arguments differ in many respects, but what they have in common is the idea of justifying a particularly basic and pervasive principle not by locating it within a certain picture of the world, but by showing it to be a precondition of drawing that picture.[11]

The manner of the Thermometer Story's failure positions us beautifully to make an argument like these on behalf of the Objectivity Principle. Scientific inquiry *just is* inquiry into how the world is, and by "the world", we mean the world that Williams has in mind, the world the Thermometer Story talks about, the world I believe in but Shmuel doesn't: a single world common to multiple perspectives occupied by different inquirers. So if the Objectivity Principle is a precondition of the proposition

[11]See Aristotle, *Metaphysics* Γ:3, 1005ff; Reichenbach (1949, ch. 11) and, for a more recent proposal in the same spirit as Reichenbach's, Schechter and Enoch (2008); Wittgenstein (1969, §163, cf. §337), and a more recent (and more cautious) deployment of the same ideas can be found in Crispin Wright (2004).

that there is such a world, then it is a precondition of scientific inquiry quite generally. But science, according to the definition I specified in the introduction, is no more and no less than our "best way of understanding the world", so we *must* be justified in pursuing it—we are doing what's best! So the Objectivity Principle is justified insofar as it is a precondition of something we must be justified in.

This argument is onto something, and my own account of the Objectivity Principle will be a kind of transcendental argument (though a unique kind). But as presented, it suffers from a fatal equivocation. We can think of scientific inquiry in two different ways. We can think of it in the normative terms I suggested in the introduction: science encompasses our *best* ways of understanding how things are. Or we can think of it in descriptive, material terms, as a particular activity involving particular methods, concepts, and axioms. These two senses must be distinguished. In particular we cannot jump off of one notion of science and onto another midstream—not if we want a valid argument. Thus we cannot offer some descriptive conception of science (it is an activity that works thus and so...), note that some principle is a precondition of science so understood, and conclude that this principle is thereby justified because science *just is* our best way of understanding how things are. But this is precisely what our transcendental argument does: it uses a descriptive notion of science in the course of the argument—science as the particular activity concerned with the sort of world the Thermometer Story talks about—but then insists its conclusion has all the force of a claim about the normative notion of science—science as our best way of figuring out how things are.

If we straighten out the argument so it is only about our normative notion of science, then it is unsound. We have been given no reasons for thinking that the Objectivity Principle must be a precondition of our best ways of figuring out how things are. If we do the opposite and make it only about the former, then it only gives us a hypothetical justification of the Objectivity Principle: a man is justified in adhering to it if he is justified in pursuing this particular conception of science. So the force of the justification is lost the moment we ask, "but is this a good conception of science—is it the one we should adhere to?".

Of course, a hypothetical justification may become categorical if its antecedent

can, for whatever reason, not be given up. But we obviously *can* give up on this conception of science. Still, maybe we can do better with a more ambitious transcendental argument, one that connects the Objectivity Principle to some ground that is inescapable in the way that our particular conception of science isn't. Because the most obvious candidate for this inescapable ground is reason, these sorts of arguments have tended to advance forms of *rationalism.*

Thomas Nagel's rationalism, to take one salient example, offers the resources to mount a defense of the Objectivity Principle with these features. According to Nagel, the thought that we live in a world common to multiple perspectives is inescapable. And once we assent to this proposition, other self-evident judgments follow. We see human beings as fully real parts of the world, and so we come to see the reasons that animate them as *objective* reasons. We see their reasons not as an impetus for them alone, but as reasons for everyone. Nagel thinks it is self-evident that your pain gives me reasons to soothe you, and that your objections give me reasons to doubt. This final thought about the objectivity of reasons gets us very close to the Objectivity Principle, for it entails that the reasons backing our theoretical judgments transcend individual perspectives. This argument is supposed to do better than our transcendental argument because reason is inescapable in a way that our particular conception of science was not. Reason is our 'last court of appeal'. It is the faculty to which we must ultimately refer all questions and challenges, even those dealing with reason itself. So we cannot step outside of reason just by asking "why be interested in this?" the way it seems we can step outside of particular conceptions of science. As a result of this status, the judgments that this faculty makes—that your objections give me a reason for doubt—have a hold on us that those connected to optional activities do not.[12]

But this advantage is a mirage, and my response to Nagel's story is similar to my

[12]I draw this reading of Nagel's program from three different books, each of which marked something of a change in view, so I do not offer it with total certainty. See Nagel (1997, p. 9) on the inescapability of reason and later in the same book (p. 92) on "the inescapability of the idea of objective reality". See Nagel (1989, pp. 159-160) for claims of "self-evidence" (compare this to a more articulated theory of rational intuition—'intellectual seemings'—developed by Laurence BonJour (1998)), and the same pages plus Nagel (1970, chh. 11-2) for claims about the objectivity of reasons. (Nagel's view on this last point has evolved in ways I cannot touch on here.)

response to the transcendental argument. It is on to a promising line of thought, but, as with the transcendental argument, it turns on a subtle equivocation—this time between two ways of thinking about reason. We can understand reason as a very basic and very slight capacity. Most activities can be challenged from the outside by asking "why should I undertake this kind of activity or care about its rules?". But, we might wonder, can we understand such challenging as a distinctive activity, the activity of asking why we should be interested in playing a game, following a principle, or believing a proposition? With this thought in mind, we might reasonably think that we cannot step outside of such an activity because it is the very thing we take up as part of stepping outside of thing. One conception of reason identifies it with this critical activity. Reason is no more than our capacity for detached reflection. But Nagel also invokes another conception of reason. We can understand reason as a robust, quasi-perceptual faculty that detects 'self-evident' facts and delivers us information about a special sort of entity, reasons.

The first of these conceptions is what offers hope for a complete and categorical justification. It is the conception that I can see turning out to be 'inescapable'. But the second is what actually delivers the Objectivity Principle. Nothing Nagel says suggests that critical reflection *per se* generates anything close to the Objectivity Principle. On the other hand, if we take the latter picture of reason, with its rational intuitions and reificiation of reasons, then we can get something of an argument for the Objectivity Principle: it is a self-evident fact that reasons are objective things, and the Objectivity Principle follows from that.[13] But relying on this conception of reason makes our justification something less than categorical, for we can just as soon give up on the judgments of this faculty of reason as we can on a particular conception of science. To this Nagel might respond that the self-evidence of these judgments makes them inescapable as well. I cannot help but think your objections give me reasons to doubt. But this is plainly a different kind of inescapability from

[13]This seems to be the tack that Nagel (1997, p. 6) takes when he says that he finds common cause with Descartes and Frege, against Hume and Kant. This suggests that Nagel fits, perhaps imperfectly, in the ranks of traditional rationalists, that is, amongst the likes of Samuel Clarke, William Wollaston, Christian Wolff, W. D. Ross, and, more recently, T. M. Scanlon (1998, pp. 17-8) and Derek Parfit (2011, ch. 2.4). Here I treat Nagel as a representative of this tradition.

the one arising from our first notion of reason: it is the sense of inescapability we associate with shackles and prisons. And just as we might think that our escape from these things is impossible but still dare to ask "but is my imprisonment just?", we can still ask "these judgments may be self-evident, but are they *right*?" And this is all we need to show that the rationalist's justification is just as incomplete our transcendental argument.

Where do we go from here? I think the arguments we have seen in this section point to a way forward, both in their insights into the problem and in the ways they fell short. On the first score, I think three points are worth holding onto: the idea that we ought to justify the Objectivity Principle by showing it to be a precondition of something else, the further thought that to make such a justification complete this something else must be somehow 'inescapable', and, finally, the suggestion that reason should play a role in all this.

On the second score, it is worth reflecting on why the particular arguments put forward proved inadequate. The way they failed has a lot to do with a hoary style of ethical argument, one whose most famous instance is G. E. Moore's Open Question argument.[14] There are several ways to read this argument, but here's mine. 'Good' is a normative term; indeed, it is our most general term of commendation. And for any state of the world, characterized descriptively, it is always an 'open question' whether that thing is good. In other words, there is no descriptive term L such that showing 'x is L' establishes, thereby, that x is good. Thus an evaluation of something as good, bad, or otherwise is always a step beyond an inventory of its descriptive properties. This argument has its limits, but Moore's basic thought is a compelling one. Establishing that a thing is of some kind (and not merely the kind 'goodness') cannot *by itself* establish whether we should commend or condemn it—that's a further question. And this is precisely the problem we found with our transcendental and rationalist arguments. If we try to justify the Objectivity Principle by showing that it is the presupposition of scientific inquiry or a deliverance of rational intuition, but

[14]Moore (1903, §§27ff).

we construe inquiry and reason as particular faculties with a particular nature, then we are free to ask what hold they should have over us, why the Objectivity Principle having a certain status with respect to inquiry or rational intuition makes it a good principle. And this leaves the justification incomplete. If we want to secure a better justification, we need a way around this pitfall.

1.4 The Objectivity Principle is the codification of the critical activity of reason

My own proposal emerges from the insights and shortcomings of the arguments from the previous section. I propose to ground the Objectivity Principle in the incredulous stare that drives the Open Question argument, in the very critical power that undermined our earlier attempts. Doing this will deliver on the insights of our transcendental and rationalist arguments. First, insofar as scientific inquiry is conditioned on this power, such an argument will be transcendental. Second, this power can be thought of as a power of reason; indeed, it is not unlike the thin notion of reason that I said was one of two Nagel equivocated between. And, third, this means that the Objectivity Principle may thereby be grounded in something properly inescapable.

But I'm getting ahead of myself. I have so far said very little about this peculiar power. The best way to understand it is to see the contrast between a creature with this power of reason and one without it. For a dumb creature, one without the power of reason, to have a certain attitude is to be automatically moved by it. When a beagle has an instinct, it is enough that it is his instinct for him to be moved by it. There is no question whether he *should* be moved by it. To say that a beagle has an attitude is just to reify his disposition to act in a certain way. But if we give our beagle the power of reason, then everything changes. Not because he can now detect special entities like the Form of the Good, but because our Rational Beagle can detach himself from his instincts, he can challenge them, and he can ask whether he should act on those instincts. In other words, the critical power of reason is the thing that opens the Open Question. It is reason so understood that enables the Rational Beagle to look down on his attitudes and say,"yes, this may be my instinct, but what should I do about

it?".[15]

I propose that the Objectivity Principle is the *codification* of the activity of this critical power. What do I mean by 'codification'? Just what I take the man who coined the word to meant by it. In 1811 Jeremy Bentham wrote to James Madison offering to codify the laws of the United States. Bentham was not offering to invent new laws of course. He was proposing to render America's existing legal structure into explicit statutes. I say the same relationship holds between the Objectivity Principle and the critical activity of reason: the former is what we get when render the natural activity of the latter *as a norm*. What the Objectivity Principle tells us to do is what the critical power of reason does by its nature.[16]

To defend this approach I must do two things. First, I need to show that my claim is true. And second, I need to show that it does not meet the same fate as our failed transcendental and rationalist justifications. I undertake these tasks in that order.

My argument that the Objectivity Principle codifies the critical power of reason is a simple one. The function of reason is to detach an agent from his attitudes and opinions for the sake of holding those attitudes and opinions up to critical scrutiny. This scrutiny, I argue, just is the exposure of those attitudes and opinions to considerations gleaned from other perspectives. And there is no principled limit as to *which* perspectives this scrutiny includes. Therefore, it is part of the critical function of reason to expose our attitudes and opinions to considerations from *all* perspectives. And this, in a nutshell, is what the Objectivity Principle requires.

This is but a thumbnail sketch of the argument; let me go through it with a little more care. The key to this argument lies in understanding what form the sort of scrutiny the critical power of reason must take. We give a beagle the power of reason and see him transform from something whose attitudes and opinions merely affect

[15]This notion of reason is similar to ones discussed by John McDowell (1996) and Christine Korsgaard (2009b). Also compare Ernest Sosa's (2009) idea of 'reflective knowledge'. The original incarnation of the idea, in modern times anyway, is due, I believe, to Kant, in particular his reply to the positive 'geometrical' conception of reason that is a forerunner of the rationalism I attributed to Nagel. I don't have space to properly discuss this here, but see the *Critique of Pure Reason*, Avi-Axii, A796/B824ff, the *Critique of Judgment*, 5:294ff, and *Answer to the Question: What is Enlightenment?*.

[16]The best example of this kind of suggestion in philosophy—that a principle codifies the requirements of a discursive activity—is the inferentialist semantics developed by Robert Brandom (1998).

him to a creature who looks down on these attitudes and opinions to ask the characteristic questions of this power: "having experience E inclines me to believe that p, but is it actually a reason to believe that p?" and "I have an instinct I to ϕ, but should I heed I and ϕ?". Some ways of responding to these questions clearly don't count as reasoning. It won't do to for our beagle to respond like this: from where I stand right now E leads me to believe that p, therefore I am going to believe that p. Nor can he say this: since, as a matter of fact, I now have an instinct I to ϕ, I am disposed to go ahead and ϕ. These episodes don't count as reasoning because they include neither of the things the critical power of reason is supposed to accomplish. There is no detachment between this Pseudo-Rational Beagle, as I will call him, and his attitudes. And so there is no reflective scrutiny. A person, or beagle, does not detach himself from his instincts if those instincts are the only things in his deliberative ledger. And it is no scrutiny of those instincts for an individual to confirm that they are, indeed, his instincts. Thus while these episodes may resemble reasoning insofar as they involve a soliloquy and the word 'therefore', they are not the genuine article because they don't have the appropriate functional profile. They don't introduce the reflective distance and scrutiny that defines reason.

What our Pseudo-Rational Beagle's case shows is that the scrutiny imposed by the critical power of reason has to involve some real friction. But where is this friction supposed to come from? It cannot come from our instincts and received opinions. That would immerse us in something like our Pseudo-Rational Beagle's rubber-stamping soliloquy. It must come from somewhere else, from something beyond our own parochial vantage point. In other words it must come from how things seem, or at least how we think would seem, *from other perspectives*.

We can reach the very same conclusion by reflecting on how the Open Question argument works. When we undermined our transcendental argument by asking "why should I be interested in this particular conception of scientific inquiry?", our question had force because we could see how things might be otherwise. Even if we are interested in this conception of scientific inquiry, we can see how we might be indifferent. Even if we think we are biologically incapable of pursuing another conception of inquiry, we can entertain the bare possibility of it. And it is awareness

of these possibilities, of imagining ourselves occupying a perspective where a particular conception of scientific inquiry does not already grip us, that gives the Open Question its force as a decisive objection rather than a idle cavil.

Of course, in practice many of these 'other perspectives' will be ones we identify with: those of our future selves, ourselves in other circumstances, our kith, kin, and countrymen. We see this in the deliberations of our Rational Beagle. He has an instinct to believe that there is a cat behind the big oak in the backyard. When the power of reason enters the picture, the beagle's instinct is no longer up to the job of bringing him to this belief. The beagle has to hold this instinct and his proposed belief up to scrutiny, and that just means examining it from perspectives beyond his instinctive self. Would he come to this belief in the absence of the instinct? Does the instinct have some special purchase on what's behind the tree?

This is not to say that the answers to these questions will trump the beagle's initial instincts. The question is one about which considerations get into the game of deliberation, what we should care about to any degree at all. And my claim is that for the critical power of reason to involve real reflective distance, and for it to be be genuinely scrutinous, considerations from other perspectives must get into the game. If we accept this connection between the critical power of reason and caring about how things seem from other perspectives, the question of the hour becomes *which* perspectives are relevant. All of them, or just a handful?

We should be skeptical that the critical power of reason can brook any circumscription at all. Suppose we have a creature, a Hidebound Beagle, who rigidly circumscribes the range of perspectives he considers to the class of other creatures just like him, to the class of Hidebound Beagles. Now because Hidebound Beagles circumscribe their deliberations in this way, there will inevitably be some judgments for which all the perspectives that they consult will agree, some judgments that are taken as gospel amongst Hidebound Beagles. These will be judgments for which their deliberations are effectively closed. So when a Hidebound Beagle tries to gain reflective distance from one of these judgments, when he tries to hold it up to the critical scrutiny of reason, he will find no friction. Because all Hidebound Beagles agree on this matter, there is not even a sensible question as to whether I should make

this judgment. Questions like "is this really a good judgment?" fizzle out because no alternative, not even the barest possibility of one, is present to the Hidebound Beagles' deliberations.

But from the outside, from our perspective looking down on the Hidebound Beagles, we can see that this question should have force. The Hidebound Beagles ought to scrutinize their judgments from perspectives beyond their own clique. And the question "should I really make this judgment?" clearly does make sense, even if they don't realize it. Moreover, from this outside perspective we would clearly fault the Hidebound Beagles with a failure of reason. They have circumscribed the range of scrutiny they are open to and thereby neutered their critical power of reason.

Now the Hidebound Beagles are clearly doing better than their cousin the Pseudo-Rational Beagle, and so we might want to introduce talk of degrees here: we can be more or less engaged by the critical power of reason. But it is clear that the Hidebound Beagles' circumscription of scrutiny amounts to a curtailment of their reasoning, even if not a total one. Equally clear is that if *we* similarly circumscribe the scrutiny to which we subject our judgments—to some class of perspectives P—we would be guilty of the same crime. For we can imagine someone standing outside of P, in some more distant perspective, and saying of us what we said of the Hidebound Beagles: that our scrutiny has been artificially restricted, that questions that appear closed to us really are open, that our reasoning is deficient because it is not adequately open-minded. And they would be right. That we can imagine such a perspective beyond any proposed field P, and be moved by the thought that this perspective matters, shows that the critical power of reason requires an openness to scrutiny from any perspective whatsoever.

Does this mean that we must be open to the astrologer, the man in the tin-foil hat, and the persistent skeptic for our critical power of reason to be fully engaged? Of course. It is nowadays taken as a sign of philosophical seriousness to insist that we ignore the loon from the get-go, but I think this betrays an alarmingly dogmatic conception of philosophy. We need not let the skeptic's arguments consume us or embark on a point-by-point refutation of astrology, but we must be open to their perspectives in the minimal sense that the Objectivity Principle requires. Being altogether deaf

to them does constitute a dogmatism contrary to the demands of reason.

Thus the critical power of reason burdens us with an open-ended sort of scrutiny, one that pushes us to take account of more and more perspectives without any natural stopping point. This a radical thesis about the nature of reason, but I think it is one we have to accept to preserve what Kant calls reason's "spontaneity". For Kant, this spontaneity consists in reason's independence from empirical conditions, in the fact that reason is not determined by any course of prior events. Kant understood this determination in causal terms, and it quickly gets wrapped up in the noumenalism of his whole moral system.[17] But I think the way we have been talking about reason so far offers better, less metaphysically recondite way to understand spontaneity. If we circumscribe the range of perspectives that are brought to bear on our deliberations, then there will inevitably be some judgments for which the outcome of those deliberations is a foregone conclusion. We saw an example of this with our Hidebound Beagles. On issues where they share a common inclination, the outcome of their 'reasoning' is determined in advance. If reason is not to be so determined, it cannot be closed off: it must be open to all perspectives in the way I require. Thus in contrast to Kant's noumenalism, I suggest that we can understand spontaneity as this special openness.[18]

With this we have established the connection between the Objectivity Principle and the critical power of reason we were after. It is the function of the critical power of reason to expose our attitudes and opinions to considerations from *all* perspectives. And this is just what the Objectivity Principle requires.

1.5 My justification defended

That brings us to our second question. Why is this justification any better than the arguments that failed in the previous section? It certainly looks like I am hanging my hat on the dynamics of a 'particular conception' of reason in just the fashion I criticized Nagel for doing. Why is the 'critical power of reason' more special than

[17]See *Groundwork*, 4:452 and the *Critique of Practical Reason*, 5:42-3.

[18]In Chapter 3 I explore this thought much more, and in particular I show how this feature gives normative discourse a special status between our traditional standards of objectivity and subjectivity.

what the transcendentalist or rationalist offers? To some parts of this charge I plead guilty. My argument *is* a kind of transcendental argument, albeit one quite different from what we saw before. And it *is* premised on reason, albeit a very different sort.

What makes my approach different from its forebears is that I ground the Objectivity Principle in the very thing that undermined those forebears, in the critical reflection that stands behind the Open Question. Why does this grant us an automatic justification? Why can't we stand outside of this activity of critical reflection and ask why it is worth our time? *Taking reflective distance* is the one activity we cannot step outside of and challenge precisely because *it is* the activity *stepping outside of*. So when we ask, "why should I engage in reason rather than something else?" we are already engaging reason. We might as well be asking, "why should I be asking this very question?". The Objectivity Principle therefore gets automatically validated the moment we open the line of questioning that might challenge it.

This riposte is very quick, I realize, so let me come at it from another angle. We have already seen that it is our ability to put reflective distance between ourselves and our judgments for that gives questions like "should I really believe this?" and "do I have a reason to believe this?" their bite. Without this power, the question is a meaningless harangue, like giving someone the raspberry. It is the critical power of reason that opens the Open Question and gives normative discourse its sense. This connection between critical reflection and normativity is familiar from the work of Christine Korsgaard. She notes that "we human animals turn our attention on to our perceptions and desires." And this special capacity for reflection, "sets us a problem no other animal has. It is the problem of the normative." She continues,"I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. [...] Now I have a problem. Shall I believe? Is this perception really a *reason* to believe?"[19] Thus it is our unique

[19]Korsgaard (1996b, p. 93). My discussion in these paragraphs is obviously indebted to Korsgaard. That said, it should be clear from the foregoing that I disagree with Korsgaard about the nature of the critical power of reason and, consequently, from where our obligation to care about other people arises (what she calls the 'publicity of reasons'). She thinks that the latter dynamic arises out of the fact that reasoning is an inherently social and interactive activity (see Korsgaard (2009b, ch. 9)). But I think this claim about reasoning betrays some residue of dogmatic rationalism. I have argued that this feature comes out of the way the reason applies critical scrutiny. I also think we can extract much less

ability to detach, reflect, and scrutinize that introduces normative questions in the first place.

This connection between critical scrutiny and normativity is an arresting one, but it is also elusive: how, exactly, does reflection generate normativity? The idea from the logical empiricists we encountered in our discussion of the Thermometer Story's circularity gives us a good way to cash this out. We have to take on certain methods, concepts, and schemes in order for certain kinds of inquiry to be possible. For instance, Michael Friedman argues that accepting Newton's laws of motion is a precondition of inquiry into the laws of gravitation. The laws of gravitation only hold in intertial reference frames, but our concept of an intertial reference is just one in which the laws of motion hold. Thus the laws of motion are a precondition of inquiry into the laws of gravitation because they help constitute an essential concept of that inquiry.[20] We can understand the connection between the critical power of reason and normativity along the same lines. It is only because we have the critical power of reason that it makes sense to ask what we *ought* believe, what we have *reason* to do, what it would be *good* to try. Thus this power of reason is partially constitutive of these core normative concepts, and so it is a precondition for normative inquiry in much the sense that Friedman outlines. (And so we can think of my justification of the Objectivity Principles as a transcendental argument relative to normativity.)

We could characterize this arrangement as a kind of 'inescapability', but we have to be careful with this word. Some things have the inescapability of shackles: they are inescapable because it is a brute fact that it is impossible to get out of them. The rationalists' 'self-evident' judgments purport to have this feature. But other things have a more essential inescapability. A pair of examples will help. Occupying space while giving a philosophical argument is inescapable, but only because of brute metaphysical facts. I have to occupy space while giving a philosophical argument, but that has nothing to do with the nature of philosophical argument, and so no one is impressed when I claim that all arguments for the ideality of space are self-defeating. On the other hand, following the rules of chess while playing chess is inescapable because

normative content from reason than Korsgaard does, something I discuss at greater length elsewhere.

[20]Friedman (2002).

that's just how chess works. And this is why the rules of chess hold a special sway over chess players. I am proposing that the critical power of reason possesses the second sort of inescapability. We are not merely dragooned into the ambit of reason. Indeed, we *can* give up on reason in some ways: we can lobotomize ourselves; we can cultivate an Übermensch-like numbness to the opinions of other people; we can jump off a bridge. The way in which reason is 'inescapable' is that it is automatically vindicated when we broach the question of its vindication—which is a normative question—just as the rules of chess get automatically vindicated when we broach questions within the domain of chess. The Objectivity Principle, insofar as it is the codification of this activity of reason, gets the same redoubtable justification.[21]

1.6 Toward the Categorical Imperative

We are justified in following the the Objectivity Principle not because there is one world common to all inquirers, nor because it is a precondition of a certain conception of scientific inquiry, not even because our faculty of rational intuition tells us so. The Objectivity Principle has much deeper roots. It codifies the activity of our critical power of reason, the very power that makes possible questions about what is justified and what we should do.

But now notice how widely this justification generalizes. The critical power of reason has just as much ability to challenge things characteristic of the ethical domain—our plans, intentions, commitments, values, our whole way of life—as it does our theoretical opinions. This is something we already saw in the previous section. The critical questions, "Should I really follow my dreams of singing on Broadway instead of staying in this sleepy little burg", "Am I justified in trusting this man in a windowless van?", and "Is material success really such a good thing?" engage reason in precisely the way that the theoretical questions I have been incanting again and again do ("Should I really adhere to the method of inference to the best explanation?", "Am I justified in trusting this spectrograph?", etc.) Thus there is no real

[21]What I am proposing here could be aptly called a version of epistemic constitutivism. See Luca Ferrero (2009) for a discussion of both the basic ideas of constitutivism and the nature of inescapability.

question of whether there is such a thing as practical reason, when we understand reason in the thin, critical sense that I am using here. It's plain as day that there is such a thing as practical reason. To modify a famous phrase from Hume, 'tis just as much reason's function to question our plan to destroy the whole world as it is to challenge our methods of inference.

This means that everything we said about how the very activity of the critical power of reason justifies the Objectivity Principle can be extended *mutatis mutandis* to a justification for an *ethical analogue* of the Objectivity Principle. This *Ethical Objectivity Principle*, we might call it, will codify reason's reckoning with our practical attitudes and opinions just as the original Objectivity Principle codified reason's scrutiny of our our theoretical attitudes and opinions. As I said in the introduction, I think this principle is a close relative of Kant's Categorical Imperative, or rather, to two formulations of it: the Formula of Humanity and the Formula of the Realm of Ends.

The Ethical Objectivity Principle will have the same two facets as our original Scientific Objectivity Principle. The first will go something like this:

> A divergence of practical opinion between individuals occupying two perspectives must represent a kind of error and is therefore a *prima facie* reason for us to correct that error, for us to restore concordance.

Let's take a few examples to see how this works. Jones thinks killing Smith is the thing to be done, but Smith has his doubts. He disagrees with the practical judgment, 'Jones killing Smith is the thing to be done'. Per the Ethical Objectivity Principle, this pair of judgments constitutes an error in need of rectification in precisely the way that a disagreement between Smith and Jones on a theoretical matter would. More generally, suppose that Smith judges something to be a worthy end. Here the Ethical Objectivity Principle entails that this judgment must hold some sway over Jones's deliberations too, that Smith's setting something as an end makes it a prima facie end for Jones as well.

With this consequence, the similarities between the Ethical Objectivity Principle and Kant's Formula of Humanity emerge. The latter is familiar enough: "Act so you

use humanity, as much in your own person as in the person of every other, always at the same time as end and never merely as means." But Kant's full understanding of this principle only emerges a few paragraphs later. Kant grants that, "humanity would be able to subsist if no one contributed to the happiness," but he insists that this would enable "only a negative and not a positive agreement with *humanity as end in itself*". To achieve the 'positive' agreement that he thinks is required, "everyone [must] aspire, as much as he can, to further the ends of others. For regarding the subject which is an end in itself: if that representation is to have its *total* effect on me, then its ends must as far as possible also be my ends."[22] And this requirement, that we regard the ends of others as holding some sway over our own deliberations, is what the Ethical Objectivity Principle also requires.

The second facet of our Ethical Objectivity Principle has much the same form as its Scientific sister:

> A total concordance of practical opinion, of judgments about what to do, what is valuable, what matters in life, operates as a regulative ideal for practical reason in just the way the analogous concordance does for theoretical reason.

This also looks Kantian to me. A few pages after expounding on the Formula of Humanity, Kant draws a corollary. From the fact that "rational beings all stand under the law that every one of them ought to treat itself and all others never merely as means, but always at the same time as end in itself" Kant says, "arises a systematic combination of rational beings through communal objective laws, i.e., a realm that, because these laws have as their aim the reference of these beings to one another as ends and means, can be called a 'Realm of Ends'." Kant goes on to say that "morality consists in the reference of all action to that legislation through which alone a Realm of Ends is possible."[23] The Realm of Ends and its "systematic unity of all ends" are very much like our state of universal practical concordance. And the fact that Kant calls this state 'only an ideal' further suggests that he intends the Formula of the

[22]*Groundwork*, 4:429-30.
[23]4:433-4.

Realm of Ends to offer a regulative ideal for the conduct of practical reason. Thus the demands of the second facet of our Ethical Objectivity Principle echo those of Kant's Formula of the Realm of Ends.

Of course, I have not said a single word about Kant's other famous formulation of the Categorical Imperative, the Universal Law formulation, nor about the Realm of Ends' variant, the Formula of Autonomy. And I have by no means claimed that the Ethical Objectivity Principle is exactly the same as the conjunction of the Humanity and Realm of Ends principles. But all that conceded, I do think it is clear that the Ethical Objectivity Principle is something *very much like* the Categorical Imperative. Both are principles require the same minimal commitment to other people occupying other perspectives—that we cannot altogether ignore their practical judgments, their ends. And both offer a loftier state as regulative ideal for our deliberations—a state in which there is a total concordance of practical opinion, a systematic unity of ends.

So where does that get us? Let's review. First, the Objectivity Principle is a very basic principle of scientific inquiry. Indeed, our difficulties in justifying it by ever more radical procedures suggests it may be the most basic principle. We found that this principle is justified because it is the codification of the activity of the critical power of reason. But then we saw that a version of this principle about ethical judgments can be justified in precisely the same way. And we have now seen that this ethical analogue is something very close to the Categorical Imperative. All this points to the thought that I began with. If we generalize the Ethical and Scientific Objectivity Principles we will reach a *Generalized Objectivity Principle* that codifies the activity of critical reason when brought to bear on *all manner of* judgments which submit to scrutiny—including, but not limited to, theoretical and practical judgments. And this Generalized Objectivity Principle will be something very like the Categorical Imperative similarly generalized. So the claim that some readers attribute to Kant, that the Categorical Imperative is the supreme principle of *all* reason, turns out to be right.

1.7 Conclusion

How does all this add up to the grand thesis that I started out with, the unity of reason? A common thought going in to these questions is that the pivotal methods of science of science and ethics, the methods that make theoretical and practical reasoning what they are, will be things that they inherit from their subject matter. And as a result these things will be rather different for science and ethics. The best example of this is the thought, which we see in Williams, is that science is characterized by a drive toward objectivity, and this drive is imposed science's subject matter—by the fact that there is a single world that our investigations are about. But ethics, according to this line of though, does not have such a subject matter and so is not driven toward the same objectivity. Many have disagreed with Williams of course, but they have happily accepted his background assumption about *why* science aims for objectivity.

By contrast, I have argued that our common assumption about why science strives for objectivity is wrong. It is not because of science's subject matter that it aims for convergence. We are driven toward objectivity by the demands of the critical power of reason. And this critical power governs not just theoretical judgments, but any judgment that we can hold up to reflective scrutiny—including judgments about how to live and what matters. And with this we get the three forms of unity I promised. Our epistemic and practical methods bottom out in the activity of a single faculty of critical reason. This faculty generates both epistemic and practical normativity. And this activity engenders a substantive norm, a Generalized Objectivity Principle, that serves as a common methodological foundation for both science and ethics.

Chapter 2

Laws of nature and laws of freedom

Constitutivism is the thesis that the solution to the problem of categorical normativity—what these we ought to do independent of our interests and aims—emerges from the very way that problem is introduced. The categorical norms of action *just are* those principles we must obey if we are to be the sorts of creatures who can perform actions. They are the 'constitutive requirements' of agency.[1]

For constitutivists, this is not an idle elegance. They think constitutivism is the only way to solve a problem with the very idea of categoricity. We think that *n* is a norm that commands categorically, that everyone ought to heed *n* no matter his plans and projects, desires and dispositions. But Jones doesn't see things this way. He doesn't see why *he* ought to heed *n* because he is dead set against what *n* requires. We explain to Jones that he should obey *n* because of some fact about that norm, that *n* is *F*. But Jones protests that we have only given him hypothetical advice: one must follow *n* if he cares about adhering to norms that are *F*. But, Jones explains, he doesn't care about *F*-style norms any more than *n* itself. And so he remains unmoved.

We might call this the problem of collapsing the hypothetical imperative. Any argument we give that one ought to follow a norm *n* can be understood by someone sufficiently incredulous or critical as having only hypothetical force. So if we want to vindicate categorical norms, then we need to find a way to collapse one of these

[1]Defenders of one form of this view or another include Christine Korsgaard (2009b), David Velleman (2009) and (2000), Connie Rosati (2003), Tamar Schapiro (2001), and Peter Railton (1997), among others.

hypothetical imperatives into a categorical one. We need to find some special F that Jones cannot deflect.

There are two models for this search. Both involve obnoxious children. The first begins with the familiar thought that our explanations of why the world is the way it is have to stop somewhere. A child asks, "why do cicadas have such an odd life cycle?" and we give an answer, "because q" (something about avoiding predators), to which the child rejoins, "but why q?," and we answer "because r." This line of questioning has to stop somewhere; there must come a time when our spade is turned. So when the child asks, "why z?" we answer, "because it's just in the nature of things that z; that's just the way it is." The normative realist wants to adapt this feature of scientific explanation to the work of collapsing the hypothetical imperative: at a certain point we just have to insist that it is in the nature of things that we are categorically obligated to n, that it's just the way things are, that our obligation is a bedrock *law of nature*.

The constitutivist, by contrast, looks to the model of the silly child who questions the constitutive aims or rules of a game. The child asks, "why should I castle now?" and we say, "because then your Rook will be free to attack my open Queen side." The child may ask further questions about the wisdom of this tactic, but if he comes to the question, "why should I try to checkmate your King?", then he has gone too far. Mating your opponent's king just is the aim of chess, so within the realm of chess strategy it doesn't make sense to ask why you should try to mate your opponent's King. Now, the child may sensibly ask his question as a way of stepping outside of the game of chess. He may wonder why the rules are what they are, or why I am forcing him to play chess in the first place. This is where the constitutivist makes her move. She says that agency is different from local engagements like chess. Agency stands to *all action* as the game of chess stands to the moves of chess, and so, unlike chess, we cannot stand outside of agency and question whether we should play the game. In other words, the constitutivist holds that agency is the widest possible such engagement, one that cannot be escaped by standing back and deciding whether or not to take it up. And from this she concludes that hypothetical norms about the demands of agency collapse into categorical norms.

So we have two ideas for collapsing the hypothetical imperative. But the consti-

tutivist thinks she has a decisive advantage. In point of fact, our obligation to, say, keep our promises is not like the fact that cicadas have an unusual life cycle: one is normative and the other is not. It makes sense for the life cycle of cicadas to ultimately come down to brute facts about particular parts of the universe being arranged thus and so. But it does not make sense for our obligation to keep our promises to come down to the furniture of the universe having a certain arrangement. It would be a very queer entity indeed that holds sway over my will just by dint of existing. On the other hand, by grounding categorical norms in the nature of agency, the constitutivist locates the seat of normativity in something that we should expect to generate norms. Our ability to deliberate and act are the very things that introduced the problem that norms are supposed to solve. So it is no more surprising that the requirements of agency hold sway over our deliberations about promise-keeping than that the rules of chess hold sway over our deliberations about whether to castle.

I am skating over some weighty issues here. I am assuming that normative phenomena really are different from cicada phenomena in a way that leaves the realist's answer to these questions wanting. And I am assuming that agency really is an engagement of maximal scope in the way constitutivists suppose it is.[2] I dilate on the point because I think it is constitutivism's *raison d'être*. The realist's attempt to collapse the hypothetical imperative requires us to posit a brute, inexplicable power that enables some select entities to reach out and command our will quite independently of how we feel about them. The constitutivist sees this and responds by proposing that we build normativity out of the activity that introduces normative reflection in the first place, out of the demands of agency.

But here's the rub. If he's not careful, the constitutivist can end up with his own, homegrown queerness. In fact, the granddaddy of all constitutivisms, Kant's, is guilty of just that. The problem is something like this. Kant's argument that the moral law is constitutive of agency depends on the will's total independence from the physical world. Here is Kant in the second *Critique*:

> Since the matter of a practical law [...] can never be given otherwise than

[2]The first point will inevitably lead back to questions of internalism. On the second point see the debate between David Enoch (2006) and Luca Ferrero (2009).

> empirically, whereas a free will, as independent of empirical conditions (i.e., conditions belonging to the sensible world), must nevertheless be determinable, a free will must find a determining ground in the law but independently of the *matter* of the law. But, besides the matter of the law, nothing further is contained in it than the law-giving form. The law-giving form, insofar as this is contained in the maxim, is therefore the only thing that can constitute a determining ground of the will.[3]

What Kant is getting at here is put well by Christine Korsgaard:

> The problem faced by the free will is this: the will must have a law, but because the will is free [...] nothing determines what that law must be. *All that it has to be is a law.*[4]

And this gets us a hop, skip, and jump away from the first formulation of the Categorical Imperative, the Formula of Universal Law. So for Kant it is *because* the will is 'independent' of the empirical world that it is governed by pure lawfulness in the way the Formula of Universal Law requires. So it is this thorough-going independence that gets Kant his constitutivism.[5]

But I don't think we can live with this kind of chasm between the will and the world. In liberating the will from the empirical world, Kant also robs it of one of its principal abilities, the ability to *affect* that empirical world. Now Kant would deny this of course. He thinks our decisions do affect the empirical world. But this depends upon the inscrutable mechanism that Kant calls noumenal affection. Our will, along with all other intelligences, configures the intelligible world, and this world, in

[3]*Ak.* 5:29. Quotations from Kant are lightly emended quotations from the following editions: *Critique of Pure Reason*, trans. Paul Guyer and Allen Wood, Cambridge University Press; *Groundwork for the Metaphysics of Morals*, trans. Allen Wood, Yale University Press; *Critique of Practical Reason*, trans. Mary Gregor, Cambridge University Press.

[4]Korsgaard (1996b, p. 98).

[5]Kant worried that without this noumenal conception of the will, agency would dissolve into the causal nexus of the physical world and the 'laws of agency' would collapse into special cases of the laws of physics. I think it is this worry, and not a bald one about libertarianism, that Kant is voicing when he insists we do better than the 'freedom of a turnspit'. Whatever our reservations about the Formula of Universal Law, it certainly seems a better candidate for a categorical imperative than the Coulomb's Law.

turn, undergirds the empirical world in the mysterious way that things in themselves undergird appearances. This 'two world' account of how my decisions manage to make a decision in the order of things seems queer to me. The realist was forced to posit a mysterious normative power that some objects hold over the will, but now Kant's constitutivism forces him to posit an equally mysterious power by which the noumenal will influences the world. This is the other side of the queerness coin.[6]

There are other strains of constitutivism, ones far from Kant's, but instead of painstakingly developing this objection against them one by one I want stop short and draw a kind of Goldilocks moral. If we try to ground normativity in something that is too much of the descriptive, physical world, then we get the realist's problem with queerness: we can explain the force of normativity only by invoking a mysterious brute normative power, the kind of thing that Mackie correctly ridicules. If we take the opposite tack and ground it in a conception of the will that is very much apart from the world, then we get Kant's problem with queerness: we can explain the will's ability to influence the world only with some equally mysterious power, something like noumenal affection.

My answer to this problem is to find a middle ground between these two extremes. I develop a 'one world' constitutivism that locates normativity in the requirements of agency, but at the same time sees agency as something that is distinctly a part of the physical world. Doing this has the welcome effect of bringing together two views that have been traditionally regarded as foes. On the one hand is the combination of constructivism and constitutivism of which Christine Korsgaard's work

[6]No modern constitutivist will talk quite the way Kant does. Among ethicists, the 'two world' interpretation of transcendental idealism has been displaced by a dualism of standpoints, between a theoretical standpoint and a practical standpoint. I find this view unsatisfactory as an interpretation of Kant. Firstly, the place where Kant talks most explicitly about standpoints is in the failed 'deduction' of third book of the *Groundwork* (see 4:459). The 'fact of reason' argument of the second *Critique*, which aims to secure a similar conclusion, is harder to make out as a claim about standpoints. Second, it is important to Kant's system that the noumenal realm is a single unity. But much of what Kant consigns to this world in the Transcendental Dialectic of the first *Critique* cannot be understood as simply how things seem when we take up a 'practical standpoint' on the world. Points of Kant interpretation aside, the problem of how the will affects the world afflicts those prefer who prefer the more innocuous dualism. There may not be a metaphysical problem of how noumenal stuff affects physical stuff, but the more fundamental question of how our practical thought manifests itself in the theoretical world endures.

is a paradigm. On the other is hand is normative realism, especially the tradition of understanding normative principles as laws of nature. My position allows us to say that the norms we are required to follow are laws of nature—just as realists do. But it also lets us say that these laws of nature are laws of agency, as the constitutivist does, and that their content is, in part, constructed out of the behavior of actual human beings. The key to combining these two things is the special reciprocal relationship between our understanding of human behavior and our performance of the actions that constitute that behavior. Because we are trying to understand something that is jointly created by ourselves and other actors, the usual distinctions that separate practical principles and theoretical claims about nature collapse.

2.1 Guiding thoughts

The key to this project is finding a way for agency to be a part of the physical world, without it thereby dissolving into the background noise of cause and effect. And the way to do, this, I propose, is to understand agency as an emergent property and its constitutive requirements as the laws of the special science concerned with that property. But let me back up and getting a running start on this idea by citing three hoary lessons about agency.

The first comes from Harry Frankfurt. Imagine a man who has been surgically tampered with in such a way that in the event that he fails to perform a particular action, an interloper can activate a device that forces him to perform that very action. Now imagine our man goes ahead and performs this action without interference. Had he not done this, our interloper would have interceded and forced him to. So in a certain sense our man had to do this thing. But, at the same time, it seems that he did it of his own free will. The lesson of this case is that freedom should not be cashed out in terms of counterfactuals about what a person could and could not have done, but as a certain *capacity* for manifesting one's preferences, values, and choices. Capacities can be undercut of course: when our interloper interlopes, our man's capacity is lost. But just because this capacity can be taken way doesn't mean it was never there. Since

our man was left alone, he was an agent at the moment of his action.[7]

The second lesson is from P. F. Strawson. If I am struck by a can of paint dropped from a broken shelf, I do not react by admonishing the shelf, by saying that it has a rotten character, or by promising revenge. I react by understanding how the shelf failed and devising a way to prevent future mishaps. But if a man hits me on the head with a paint can, I do react this way. I rain down epithets, and I explain his behavior by citing his long-standing hatred of me, his desire to see me suffer, and the malice that animates all his behavior. What lies behind this dichotomy of reaction is the thought that a certain kind of explanation is appropriate for inanimate objects while another kind is appropriate for people.[8]

Marrying this lesson to Frankfurt's, we can say that particular attitudes, including a particular approach to explanation, are uniquely appropriate for those possessing a special capacity, the capacity of agency. Thus a person possesses this capacity if and only if it is appropriate to explain her behavior in this particular way.

The third lesson comes from Daniel Dennett. It is that this special mode of explanation amounts to a kind of pattern recognition. We are confronted with a lot of behavior that might otherwise look like noise, but we can make sense of it by locating patterns in that behavior and ascribing particular states (beliefs and desires, say) on the basis of those patterns. The special sciences are a tonic comparison. If I try to understand the sociology of an ant colony just by looking at the microscopic particles that constitute it, I will be frustrated, for all I will see is a maelstrom of useless data. But if I look for patterns at a lower level of resolution, if I look for *ants* and *tunnels* instead of clouds of particles, I can begin to understand what is going on. And if instead of particular ants and particular tunnels I look for *ant roles* and *ant hierarchies* I can understand the sociology of the colony even better. The thought for action is similar.

[7]Frankfurt (1969). Some writers, for instance Andrews Reath (1997), see Kant's constitutivism as following from a capacity view of agency. They argue that following the moral law confers certain *normative* capacities on an agent, like the ability to hold someone responsible.

[8]Strawson (1962). One might say, and indeed sometimes this seems to be Strawson's attitude, that explanation, in general, is incompatible with the reactive stance. I don't think this can be right. Even when taking a reactive stance we need some understanding of a person's behavior; taking this stance doesn't mean treating someone like a black box. What's distinctive of this stance is the *kind* of understanding we seek.

We have all this physical data about people moving about. It is hard to make heads or tails of it just by looking at the whirring of little particles. But if we look instead for patterns that emerge at a coarser grain of resolution in all this noise, at beliefs, desires, intentions, reasons, and choices, then we may begin to predict and explain the behavior of these complex systems.[9]

If we put these three lessons together, we get a preliminary conclusion that makes constitutivism a bit easier to swallow. Free agency is a capacity we ascribe to a complex system if and only if it exhibits patterns that make it appropriate to take a particular approach to explaining that system. And if there are particular standards to which our behavior must conform for these patterns and regularities to appear, then these standards will be constitutive requirements of agency. Adherence to them would be, in effect, a prerequisite for explaining a person as a free agent, and therewith a prerequisite for free agency itself.

2.2 Interpretation as a special science

I propose that we understand the principles of agency to be principles of a *special science*. As the psychologist, evolutionary biologist, and economist recognize different patterns that emerge amidst the whir of nature, so we should notice special patterns tied to the capacity of agency.

More particularly, special sciences offer us to the two features that our Goldilocks moral demands. Biology is very much *of* the physical world: it concerns physical things made up of the stuff that physics studies. And for this reason, the results of other sciences duly affect the course of biology. For instance, the chemistry of nucleotides has a tangible effect on how we understand inheritance. But biology is no pleonasm either. The the laws of biology do not reduce to the laws of physics. The concepts of biology carve up the world in ways that no conjunction of concepts from physics can mimic. And the explanations offered by biology cannot be readily replaced.

We should want to say the same things about agency and its study, and to do this

[9]Dennett (1991). There is a vast literature, much it more rigorous, on what features of the world ground the use of higher-level sciences like these. I mention Dennett because I like his simile.

I submit that the problem of agency defines a scientific *discipline* in the sense laid out by Stephen Toulmin and a *field* in the sense of Lindley Darden and Nancy Maull.[10]

But if the study of agency is to be a special science, what kind of special science shall it be? One thought is that it should be folded into a familiar field, something like human psychology or the study of human nature more generally. I think Philippa Foot sometimes hints at this idea.[11] If we took this route, then we would say that certain generalities about human beings ('Aristotelian categoricals') correspond to categorical norms because failing to adhere to them would make an individual something less than fully human.

But this approach won't work because it promises to turn all of humanity's warts into normative ideals. It is rare to find someone who doesn't commit the conjunction fallacy (the one exemplified by Kahneman and Tversky's 'feminist bank teller' case), and indeed we might think that there is a good reason why human nature evolved to rely on the heuristic that lies behind that fallacy. But surely the conjunction fallacy is a genuine fallacy, despite its inveterateness. So it would be quite absurd to call it a categorical requirement on the basis of its being so common. What this shows is that if we are to get the kind of constitutivism I am after, it must be grounded in a special science concerned with a phenomenon narrower and more refined than generalizations about human behavior.

This move is sure to arouse suspicions. We are not justified in postulating this special science on theoretical grounds, the objection goes, but are only tempted to do so only because we want to validate the lion's share of our common sense judgments about morality and rationality. So positing a special science of some phenomenon 'narrower and more refined' than human nature is an illicit bit of reverse engineering.[12]

The fairness of this objection hangs on whether there are phenomena distinctive enough to warrant the introduction of this extra special explanatory apparatus, whether this kind of explanation would explain something that our more flat-footed

[10]See Toulmin (1972, pp. 141-5) and Darden and Maull (1977).

[11]See her argument for how a naturalist can live up to 'Hume's practicality requirement' in Foot (2003, ch. 1).

[12]Thanks to David Enoch for posing an objection like this one.

study of human nature cannot. But there is a special class of phenomena that warrant this extra mode of explanation: our ability to control our actions by reasoning. It may be altogether normal to be tripped up by the conjunction fallacy, but we are not doomed to see this fallacy as an essential part of our nature. We can view it as a defect, and we can call it as much when we see it in other people, because we have the ability to detach ourselves from our tendency to make this conjunction, and hold it up to scrutiny—just as we can with our other attitudes and instincts. This ability is a power of reason, and it allows us to transcend the part of our nature that leads us into the fallacy, and to expect the same elevation in other people. Thus the phenomenon that requires some explanatory approach distinct from the one offered by Foot and company is our ability to regulate our behavior through this power of detachment and critical reflection. And we see this special approach to explanation on display in our ordinary interactions. When a friend commits the conjunction fallacy, we do not excuse her on the grounds that the fallacy is 'normal'; instead we think she is open to criticism because we understand her as the sort of creature who is generally capable of the reflective scrutiny that allows her to overcome the fallacy.

I don't have a good name for this special explanandum—not beyond 'agency' anyway—but I do think there is an obvious candidate for the kind of explanation appropriate to it: the method of interpretation. William Child gives a limpid thumbnail sketch of this special method.

> Interpretation is the process of ascribing attitudes to an individual on the basis of what she says and does. When we interpret someone, we aim to make sense of her by attributing beliefs, desires, intentions, emotions, and other propositional attitudes—attitudes in the light of which her behavior is intelligible as, more or less, rational action. Interpretationists think that we can gain an understanding of the nature of the mental by reflecting on the nature of interpretation.[13]

This is a good description of the problem, but it may make the method of interpre-

[13]See Child (1996, p. 7). The method of interpretation is developed in slightly different forms by Donald Davidson (sprinkled throughout the essays in his (1980) and (1984)), David Lewis (1974), and Daniel Dennett (1987).

tation sound more fixed than it really is. A more schematic, but also more accurate picture of the problem is this. We have these phenomena: human behavior guided, in one way or another, by reflective reasoning. But many different theories of human behavior are consistent with these phenomena, and so to overcome this underdetermination we have to introduce constraints on our explanations, and the natural place to look for these constraints is folk psychology. We introduce new constraints until we are able to winnow down the class of explanatorily adequate theories to just a single account. Thus the game is to find the most plausible constraints that yield a single theory. In David Lewis's words, "if you ever prove to me that all the constraints we have yet found could permit two perfect solutions [...] then you will have proved that we have not yet found all the constraints."[14] So the suggestion that our interpretation should be cashed out in terms of beliefs, desires, and intentions, and that we should assume our subjects adhere to rules like utility maximization and conditionalization are not gospel. They are good candidates for the constraints because they give us good explanations, but they are in no way definitive of interpretation.

We should therefore think of the method of interpretation as encompassing whatever helps us best explain the phenomena related to agency. *But*, at the same time, we should think of agency as the thing that interpretation is concerned with explaining. This may sound like a tight circle, but it is not a vicious one. Neither our idea of agency nor the methods we use to understand it are fixed in advance. Instead we begin with an inkling of both. Agency is something like our ability to regulate our behavior through the reflective powers of reason (and other powers that this power helps us cultivate) and interpretation is something close to our ordinary ways of understanding our friends and neighbors. Then we see the two evolve symbiotically. This is not an unusual arrangement. We see the same co-evolution of explanation and explanandum in other special sciences, for instance in our understanding of genetics and inheritance.

The particulars of the method of interpretation will matter very little going for-

[14]Lewis (1974, p. 343). I go on later in to disagree with Lewis's commitment to a single account. I think a certain amount of indeterminacy is unavoidable, and it results in a modest kind of normative relativism. For now this departure from Lewis doesn't matter.

ward. What matters more is the boundary between interpretation and other explanatory strategies. Interpretation is one way to explain the behavior of human beings, but it is not the only one. Sometimes attributing beliefs, desires, and episodes of reasoning to a person (or whatever interpretation ends up consisting in) will not help us understand his behavior—or at least not provide the *best* explanation of that behavior. When this happens we are obliged to look elsewhere. We should offer an explanation that comes out of a field other than interpretation, one that proceeds by different methods, uses different concepts, and adverts to different laws. This is precisely what happens in some familiar cases of scientific explanation: when we see that some phenomenon resists the kind of explanation typical of classical genetics, we resort to a somewhat less elegant, but perhaps more empirically adequate, biochemical approach. I will call the union of all these other explanatory approaches *mechanical explanation*. The quintessential mechanical explanations are ones that make sense of someone's behavior by pointing to an external force (like being clutched by another person or struck by a boulder), an internal force (a searing headache), or states like being petrified by anxiety, numbed by tedium, and delirious with joy. These paradigms can be misleading. Mechanical explanations are not all about emotion. Saying that someone is obsessed with the performance of a certain syllogism is just as much a mechanical claim as saying that he is overcome with sorrow, and in many cases the exhibition of particular emotions is precisely what makes a person intelligible. Moreover, not all mechanical explanations are clunky and hydraulic. Mechanical explanations are anything that attempts to assimilate a person's behavior to some explanatory rubric other than the one employed by interpretation. And there are plenty of these, many of which are nothing like being struck by a boulder. Our mechanical methods can offer an explanation of a person's behavior akin to our explanations of reciprocal grooming habits of chimpanzees, the instinctive reactions of a moray eel, the sexual reproduction of liverworts, the circuitry of a motherboard, the plasmolytic reactions of fern cells, the electrical potential in a twelve-volt battery, the mechanisms of a Swiss watch, or the collision of two billiard balls. These are all different kinds of explanation belonging to different scientific fields or disciplines (in Toulmin, and Darden and Maull's sense of those words)—primatology, electrochemistry, ballistics, watch-

making, and so on. Each of these fields has the potential to be adapted to offering an explanation of certain kinds of human behavior, sometimes even a better explanation than interpretation.[15]

My characterization of interpretation and mechanical explanation makes the line between them blurry. But this is as it should be. In most interesting cases we will use both methods together and will not pay much attention to the place where our explanations stop being interpretative and become mechanical. It is hard to tell whether a person is experiencing an emotion because he is overcome by it or because he determined that it is the appropriate thing to experience—whether he has decided that crying is the way to express the sorrow he feels or whether this sorrow overcomes him and he weeps uncontrollably. In these cases we hedge. We offer a story that has aspects of both mechanical explanation and interpretation. This case and the hundreds like it make it impossible to mark off where interpretation ends and mechanical explanation begins with any precision.

It follows from this that my claims about the connection between interpretation and agency must be reread to admit of degrees. The question is not whether we interpret a system, but how much of our best explanation is an interpretation—not whether a subject is an agent, but how much of an agent she is. This feature of my account offers the resources to respond to the familiar objection that constitutivism makes bad action impossible. If a violation of a categorical norm amounts to a deficiency in action, then how can someone actually *do* anything wrong? The beginning of an answer comes with the observation that a person's behavior must be interpreted holistically. If we do not understand a person's behavior one action at a time, but as a unified whole, then we see an action that runs contrary to the demands of agency as diminishing a person's agency without completely vitiating it. And so we can understand this action as a mistake by a genuine agent so long as that person as surpasses

[15]To anticipate any confusion, I should say that I do not see interpretation as being as radically different from other forms of explanation as some do. Many advocates of the method of interpretation, for instance Charles Taylor (1971), see it as an alternative to explanations that invoke traditional notions of causation, regularity, and prediction. I don't. Indeed I doubt that something can count as explanation if it eschews these categories too aggressively. What distinguishes interpretation on my view is the concepts it invokes and the kinds of regularities it traffics in, not a total abandonment of the contours of explanation in general.

some minimum threshold of agency on the strength of his other, presumably intelligible, actions. Thus when an otherwise normal person violates a norm we can understand as a bad action of an imperfect agent, rather than something automatically beyond the pale of agency. Most of our evaluation of actions as bad or good will therefore take place in the vast realm between minimum agency as sage-like proficiency. More needs to be said about this kind of defense, but it's not the focus of this paper, so I'll leave it there.

We have big three ideas on the table. First, a subject is a free agent if and only if it is appropriate to explain his behavior in a distinctive way. Second, this distinctive method of explanation is interpretation. And third, this connection is not all or nothing, but a matter of degree. Putting these three ideas together yields our first lemma:

> *Lemma 1.* The extent to which a subject is a free agent is proportional to the extent to which it is appropriate to use interpretation to explain her behavior.

To derive my first thesis I will combine this lemma with an account of the form that interpretation must take.

2.3 Subsumption under laws

I say that interpretation is a special science like evolutionary biology or economics. It is the special science uniquely suited to explaining complex systems that behave as agents. The next step in my march toward constitutivism is to connect this thesis to a point about laws. Consider the following three examples of special science laws.

1. The law of allopatric speciation relates the concepts *species* and *selection pressure*: Isolated populations undergo genotypic or phenotypic speciation as they become subjected to differential selection pressures.

2. Carnot's rule relates the concepts *engine*, *efficiency*, and *Carnot engine*: No engine operating between two heat reservoirs can be more efficient than a Carnot engine operating between the same reservoirs.

3. Okun's law relates the concepts *gross domestic product* and *unemployment rate*: The difference between actual and capacity gross domestic product is proportional to the difference between actual unemployment rate and the rate at full employment.

Special sciences like evolutionary biology, thermodynamics, and labor economics carry their explanatory burden by showing that their target phenomena instantiate regularities, or laws, like these. What makes the explanation of the special sciences truly indispensable is that these laws cannot be reduced to laws of more basic sciences without loss of explanatory power. If we try to render the regularities mentioned above as a great constellation of laws of physics, we lose some understanding.

This is a corollary of the much more general thought that all explanation is a matter of subsuming some target phenomena under laws. The most familiar enunciation of this idea is Carl Hempel and Paul Oppenheim's covering law model, according to which an explanation is a deductive argument that our target phenomena follow from universal laws and initial conditions.[16] The letter of the covering law model probably founders on now-familiar counterexamples, but the spirit of it is alive and well in the idea that explanation is a matter of showing that a phenomenon 'makes sense' insofar as it is part and parcel of some broader system of regularities.[17] If we understand the subsumption under laws in a broader way, as showing how these phenomena are part of larger regularities, then I think the claim that the special sciences explain by subsuming phenomena under special laws becomes rather uncontroversial.[18]

We should expect matters to be no different when it comes to the special science of agency. So if I am right that interpretation is the preferred method of a kind of

[16]Hempel and Oppenheim (1948).

[17]For an enunciation of this idea shorn of Hempel and Oppenheim's positivism see Michael Friedman (1974).

[18]James Woodward (2000) argues that explanation in the special sciences involves invariance relations and *not* laws. I think the very weak understanding of laws I am offering here manages to accommodate even his view, which is the most anomic view of the special sciences I know of. There are of course some philosophers of science who suggest we do without laws altogether, Bas van Fraassen and Ronald Giere most prominent amongst them. I think these proposals are best understood as advancing especially light-weight accounts of what laws might be, rather than saying we do away with laws altogether.

special science, then it too must have distinctive laws. To offer an explanation by interpretation is to show how someone's behavior fits a pattern—in other words, how it conforms to a special law of interpretation. This connection gives us our second lemma:

> *Lemma 2.* To explain an individual's behavior using the method of interpretation is to understand it as adhering to special laws of interpretation.

Putting our two lemmas together we get our first major conclusion:

> *Generic Constitutivism.* The extent to which a subject is a free agent is proportional to the extent to which we can understand its behavior as adhering to special laws of interpretation.

This thesis is little more than an explication of the lessons we reviewed above. Agency is intimately connected to how we react to and explain a subject's behavior. What this ultimately amounts to, when we understand some basic lessons about explanation, is the relationship enshrined in Generic Constitutivism.

We can understand this thesis on the model of what some philosophers of science call 'meta-laws'.[19] Meta-laws do not explain our target phenomena directly, but instead place constraints on what our explanations may look like. They are laws about laws. This role may seem pleonastic—why do we need a nomic level above the first-order laws of nature? Well, the laws of nature could have been radically different, and so there are many different possible sets of laws. We can think of science as the business of locating the set of laws that is actual. One way of winnowing down the set of possible laws is to notice that only some of these laws will obey what physicists call 'symmetry principles', for instance the principle that laws are invariant under spatial transformations. If we endorse this symmetry principle, we have in effect endorsed a meta-law about nature, a law about which laws are admissible, and this allows us to exclude a lot of candidates from the get-go.[20]

[19]See, e.g., Marc Lange (2009, §3.4).

[20]This view can actually be read into Kant's own philosophy. In the introduction to the first *Critique* (at B18) he offers examples of synthetic *a priori* propositions from the successful parts of science.

Meta-laws about agency will play just this role in interpretation. There are many different ways that an agent might behave and many different principles that her behavior might instantiate. What our meta-law says is that these principles must be proper laws if they are to explain anything. If we cannot explicate our subject's behavior using laws of interpretation, then we have to turn to some other approach that can. And this means that our subject is that much less of an agent. Some physicists have suggested that all *a priori* reasoning in physics involves laying down meta-laws like these.[21] In particular, the thought is that meta-laws codify methodological prerequisites: what we must assume in order to undertake the inquiry that might result in the discovery of particular laws. Thus these meta-laws are *a priori* insofar as they are preconditions of a particular kind of inquiry. The most familiar examples of this kind of proposal come from physics, where some have argued that symmetry principles represent nothing more than our need to assume that the universe is invariant across certain changes in perspective, which is the essence of our assumption of objectivity.[22] Generic Constitutivism has the same status. It is a meta-law derived just from the requirements imposed by what an explanation of behavior must look like. This thought puts in stark relief the differences between my Generic Constitutivism and Kant's argument for his version of the same. Kant got his result by claiming that the truly free will must be governed by a law with no empirical content whatsoever, an insistence that leaves us with a 'law' that requires nothing more than lawfulness itself. But we can get the same result more modestly by exploiting some ideas from the philosophy of science. The science of agency is an autonomous field whose principles

Amongst these is a conservation principle for matter. Kant's claim—if we project my vocabulary on him—is that this conservation principle is a meta-law of nature because we must assume such conservation for experience of nature to be possible. One way that modern philosophers of science who mine this Kantian vein have improved on Kant's own account is by tying such meta-laws not to experience, but to a more nebulous notion of inquiry. For example, according to the physicist Eugene Wigner (1967, p. 29), invariance principles are a prerequisite for the very possibility of discerning the laws of nature because "if the correlations between events changed from day to day, and would be different for different points of space, it would be impossible to discover them."

[21]For instance Hermann Weyl (1952, p. 126): "As far as I can see, all *a priori* statements in physics have their origin in symmetry."

[22]Again, Weyl is the most forthright, saying that "objectivity means invariance with respect to the group of automorphisms."

are not reducible to those other sciences and, like all fields, its methodology will be guided by meta-laws. And these entail Generic Constitutivism.

Let me sum up what I have said so far by showing how it suggests a reply to the man like Jones, one who rejects a particular norm *n*. We can say to Jones: insofar as you have any ends at all—as far as you aim to do anything—you must be the sort of creature that can consider ends, reason about means, deal with unforeseen circumstances, and so on. But you are such a creature only insofar as you are an agent, and so you are an agent only insofar as you adhere to the laws of interpretation. Amongst these laws is *n*. Therefore, should you violate *n*, you will thereby be that much less interpretable, and so that much less of an agent, and so that much less of the sort of creature who can have ends. This is what makes *n* categorical.[23]

[23]This point is as good as any to pause for some product differentiation. Recent years have seen two other writers plumb the relationship between interpretability and ethics. The first is David Velleman (in the essays collected in his (2000) and (2006), and then with some revisions in his (2009)). Velleman holds that the constitutive aim of action is self-understanding. A woman's deliberations are correct just insofar as they produce actions that are intelligible to her. My constitutivism starts off very differently from Velleman's. First, I do not propose any constitutive *aim* for action; I say that the laws of interpretation are constitutive *requirements* of action insofar as an action has to be interpretable for it to *be an action*. Saying that something is the aim of action suggests that it is not a structural prerequisite for something to count as an action, but an altogether stronger thesis about the *telos* of action. This difference should matter to naturalists. It is one thing to say that being thus and so is a constitutive requirement of being a chromosome or a galaxy or an agent; we find this kind of claim in ordinary scientific reasoning. It is quite another to say that having such and such an *aim* is a constituent of being a chromosome or a galaxy or an agent. It is harder to find evidence of this in ordinary inquiry. A second difference is that I offer no particular candidate for what these requirements are. I show how there could be such requirements. Velleman, by contrast, steps up with a very precise proposal: self-understanding is the constitutive aim of action. I find it very dubious that there is a single constitutive aim that we can state so succinctly and arrive at through *a priori* reflection. I think Velleman comes to this view because he is seduced, as Aristotle was, by analogies between action and belief. Velleman thinks belief has an obvious and *a priori* graspable aim—truth—and is interested in finding the analogous aim for action. But this claim about belief is false. There are lots of ways that belief can be good and bad. And the thought is even more suspect when extended to action. Sartre surely had it right when he said that human beings are not like paper knives created for a particular function. (The whole game of constitutivism, as I understand it, is to find some account of the nature of action that can produce some informative categorical norms while still holding onto this basic point—without relying on a cartoonish reduction of action.) Finally, Velleman is interested in our interpretability only insofar as it enhances our *self*-interpretability. Being interpretable to others, he says, helps us engage in repeatable scenarios that make our behavior more intelligible to ourselves, and so it is worthwhile for just that reason. In a similar vein, Adam Morton (2003) emphasizes the importance of interpretation to coordination problems between agents in which the best solution for all agents is only attainable if they

2.4 An objection: Inscrutability for fun and profit

But isn't this argument undermined by all the people who are not only difficult to interpret, but actually cultivate inscrutability—and do so to great success?

The songstress Lady Gaga once wore a dress made of meat. In response to a question about his advice to young people, Bob Dylan produced a light bulb from his pocket and responded, "keep a good head and always carry a light bulb." When signing autographs for fans, Salvador Dalí would refuse to return their pens. When interviewed on American television he referred to himself only in the third person and remarked, in a matter-of-fact tone, "Dalí is immortal and will not die." On a different television appearance he brought with him a model rhinoceros and insisted on sitting on it instead of the chairs provided.[24]

Gaga, Dylan, and Dalí are odd ducks, but we think of them as successful, at least by whatever standards they make for themselves. And so they seem to represent a problem for any theory connecting normativity and interpretability. I say that the laws of interpretation are categorical norms because adhering to them is required for agency. So when I meet someone who violates those norms, as our eccentric trio do, I should say to them, "look, insofar want to do anything at all—from wearing a suit made of meat or sitting on a rhinoceros—you have to be an agent, so you really must follow these principles." This appeal seems feeble in the face of the fact that Gaga, Dylan, and Dalí *are* successful, very much *because* they violate the demands of interpretability.

But these examples can be misleading. Our three oddballs behave strangely enough to render the heuristics we use to deal with most behavior ineffective. When we see Smith wearing linen in July, we infer that she is doing so because she wants to be comfortable in the heat, but this kind of reasoning just won't work when we encounter Gaga and her meat dress. And this failure of our superficial methods of interpretation

---

can understand each other. I see a much more direct connection between interpretability and agency. Interpretability is not merely an aid to some further goals of practical reason (self-understanding or utility maximization); it is part of what it takes to *be* an agent.

[24]Thanks to the many questioners in Madison who put the problem of Lady Gaga to me so forcefully.

leaves us flummoxed. But our initial bemusement does not mean the total and utter failure of interpretation. We have enough data about Gaga to retrench and give an explanation of her behavior that does not rely on heuristics about the relationship between season and fabric, a an explanation that starts from a more fundamental place. One obvious interpretation sees Gaga with a jaundiced eye. Gaga wants fame and fortune, and she has decided that frustrating our day-to-day expectations is a good way to do it. The public likes the bizarre and unpredictable, and they will pay good money to see it. So Gaga chooses to wear a meat dress because it will drop jaws and raise eyebrows. And this, through mechanisms all too familiar, will bring her fame and fortune. A more subtle interpretation sees Gaga's behavior as an artistic challenge to conventions that make our usual heuristics effective. As Marcel Duchamp put a urinal in an art museum to upset our bourgeois expectations about what we might see there, so Gaga puts on a meat dress to upset our expectations about what a popular singer wears. Thus her goal is didactic: to shake up the public and allow them to see their preconceived notions as optional. Whether or not these interpretations reveal Gaga to be a genuine agent, the important point is that despite its superficial intractability, we can make a start at understanding her behavior.

Indeed, it seems that the whole force of this objection, that Gaga would be immune to my appeal because she is *successful* despite her uninterpretability, assumes that we can make sense of what she is up to enough to call her successful. We might call this the self-effacement of inscrutability. Just as trying very hard to be happy may actually impede your ability to be happy, persistently aiming for inscrutability can make you an open book. Think of the wearisome comedian who offers an unending string of absurdities in lieu of wit, the philosopher who jokes, "I do everything I can to be *difficult to understand*,"[25] or one's third hour in a Dadaist exhibition.

A good example of this predicament can be found in the tricky business of creating a character that is genuinely difficult for a reader to understand. The eponymous narrator of Italo Svevo's *Zeno's Conscience* offers a good case study. As James Wood explains, it takes careful management of Zeno's reports by an intelligent and orderly author to avoid the tedious inscrutability of the Dadaist.

[25]Nietzsche, *Beyond Good and Evil*, §27.

> Zeno's narration is as fantastic as his mind, and he is therefore a highly unreliable narrator, just as Quixote would be were he telling his own tale. In most novels, unreliable narrators tend to become a little predictable, because they have to be reliably unreliable. Their unreliability is manipulated by the author: indeed, without the writer's reliability we would not be able to 'read' the narrator's unreliability. It is true that, after a few pages, we learn to discount Zeno's claims for himself; we learn to believe almost the opposite of what he tells us. This offers us, in part, the comic prospect of the patient 'resisting' our diagnosis: we, the readers, become Zeno's analysts. So the more Zeno tells us he is strong, the more weak he seems. The more he tells us that he will give up smoking, or his mistress, the less likely we are to trust him. The more he fixates on an organic cause for his many illnesses, the more we take him to be an obvious example of a *malade imaginaire*.[26]

Gaga, Dylan, Dalí, and even characters written to be incomprehensible generally want something: money, fame, a revolution of sensibilities, to frustrate voyeurs. It is an open question whether these ends are completely compatible with agency—I will go on to suggest they probably aren't—but once we realize that we already understand their eccentricities as part of some larger occupation, we can at least make a start on understanding their behavior. And this is enough to take the teeth out of the objection. Of course, if their eccentricities cannot be so understood—if they really are symptoms of insanity—then my reply is void. But so is the objection.

2.5 Kant's Assumption and Formula of Universal Law Constitutivism

With this objection defanged, we can move on to a thinking about just how interesting a result Generic Constitutivism is. It says something important, but, as with most 'meta-laws', it doesn't give us much specific content. In particular, if we had hopes of showing that a substantive moral theory arises out of the demands of agency, then Generic Constitutivism isn't much help.

[26]Wood (2002).

In order to overcome this problem Kant makes a momentous assumption. For reasons I won't go into, he assumes that the cardinal concept in our explanations of action, and thus in the laws of agency, is that of a *maxim*: an attitude that directs the will by connecting means and end. It is a principle we would articulate as "I will ϕ in order that p." This assumption is too radical a reduction of moral psychology to be true, and in later sections I will develop a version a version of constitutivism that does not rely on it, but we can learn some useful things if we provisionally accept it.

For a start, this assumption yields a strong version of constitutivism very quickly. Kant's Formula of Universal Law requires that you "act only on a maxim that you can at the same time will to be a universal law." Of those maxims that fail this test, some will violate the Formula of Universal Law because we cannot even *conceive* of that maxim as a universal law.[27] We have a perfect, or strict, duty to avoid these maxims. This suggests a narrower version of the Formula of Universal Law:

> *Perfect Formula of Universal Law.* Act only on a maxim that you can conceive of being a universal law.

If we assume a reasonably strong connection between conceivability and possibility, then a principle that we cannot even *conceive* of being a universal law could not possibly be an *actual* universal law. So if a person violates the FUL, then his action simply could not instantiate a law about how people act. But those laws, per our assumption, are just the ones that we must use to interpret an agent's behavior and, therefore, the laws that someone must adhere to if he is to be an agent. This gives us the connection between agency and the FUL we are looking for:

> *FUL Constitutivism.* Insofar as it is appropriate to explain an agent's action in terms of maxim adoption, adhering to the FUL is a requirement on agency.

To see what exactly this thesis comes to, we have to be a little more precise about what a maxim is and how these things get involved with laws. A maxim involves two

[27]See 4:421-4. The FUL's variant, "act as if the maxim of your action were to become by your will a universal law of nature," is equally germane.

things: setting an end (for instance, getting some ready cash or impressing a suitor), which can be analyzed as a kind of desire or, better, a very nebulous intention; and, second, choosing a means to that end. Thus we can take the adoption of a maxim to be a psychological state like so:

A intends that p & A chooses ϕ-ing as a means to p.

Talking about A's 'choosing' is too clunky for this account to be fully satisfactory, but it is good enough for our purposes. What FUL Constitutivism requires is that the connection between adopting a maxim, so understood, and the action the maxim it is meant to produce must be *lawful*. But what do we mean by lawful? If we adopt a counterfactual analysis, then the relevant law will look like this (representing our counterfactual as a subjunctive conditional):

$(\forall x)$ were x to intend that p and choose ϕ as a means p, then x would ϕ,

or, in symbols,

$(\forall x)$ x intends that p & x chooses ϕ as a means to $p \mathrel{\Box\!\!\rightarrow} x$ will ϕ.

On the standard story, a counterfactual '$A \mathrel{\Box\!\!\rightarrow} B$' is true just in case B is true in all the possible worlds sufficiently near (according to a similarity ordering introduced by context) the actual world in which A is true.[28] For example, a maxim like "I will bake a cake in order that I have something to eat" must instantiate a law, "$(\forall x)$ were x to intend that he has something to eat and choose baking a cake as a means to having something to eat, then x would bake a cake." For this law to be true everyone in nearby possible worlds who wants something to eat and chooses baking a cake as a means to that end must indeed bake a cake. If this condition is not met, then the principle will not be a law.

This counterfactual account is one way of making precise the idea that laws must represent genuine regularities rather than accidental generalities. It rules out cases

[28]See Lewis (1973). There are of course bulkier understandings of what laws are, as I mentioned when expressing some caution about the assumption that special science explanations must be driven by the assimilation of phenomena under laws. I will try to be as liberal-minded as possible in my discussion of laws.

where the connection between maxim and action is a serendipitous feature of the actual world by requiring that this connection also hold in nearby possible worlds. If the maxim-action principle does not meet this standard, then we cannot understand an agent's action as part of a larger pattern of maxims and actions, and so our would-be explanation is inadequate. For example, consider a possible world in which individuals accept our maxim's end, and choose our maxim's means, and yet do not take those means (e.g. they do not bake a cake). If this world is sufficiently nearby to be relevant to the truth of our counterfactual, then it undermines our proposed law. This means we cannot give an interpretation of this individual's behavior by ascribing such a maxim to him. Saying that he adopts such a maxim will not actually explain his behavior, since that adoption is not reliably connected to that behavior in a way that supports explanation.

This requirement of counterfactual robustness is what gives FUL Constitutivism its force, as it is here that pressure is applied by our need to accommodate our laws of agency with our other natural laws. A law of agency is true just in case it holds in the appropriate nearby possible worlds. But which worlds are nearby will depend, in part, on other natural laws and subnomic facts. So which maxim-action connections are law-like will depend on other characteristics of the world. Physiology, neuroscience, sociology, economics, and so on constrain *each other* in the explanations they give because they aim to explain different parts of a common world. This means that the laws of each domain are all mutually constrained. What we have seen is that interpretation and the laws of agency are part of this same system, and so which maxim-action connections are lawful will depend on how the world is according to all these other sciences.

To take a simple example: suppose it is a law of chemistry that whenever someone is in a yellow room they cannot bake a cake; but nothing at all stands in the way of people in yellow rooms forming maxims about cake baking. (Of course this is *not* a law of chemistry...) In this case, there will be a nearby world in which an agent is in a yellow room, adopts a maxim to bake a cake, but, because of this strange law, cannot effect the baking of the cake. And this world witnesses the falsity of our putative law connecting maxim adoption and action, which in turn, shows that an agent cannot

perform this action and be understood as acting on lawlike maxims.

We can see this all play out in one of Kant's examples of a bad maxim. Suppose MacDonald resolves: "I will make a false promise to get some ready cash". We can easily imagine people being much more venal and dishonest than they actually are, and this suggests that there are nearby possible worlds in which throngs of people choose making false promises as a means to getting some ready cash. But in a world of such pervasive deceit, it seems quite likely that the entire enterprise of promising would break down. With false promises so ubiquitous no one would take another's word as a guarantee. So one could not actually succeed in making a false promise, or any promise at all. One might utter the words "I promise to pay you back", but there would be no recognition of this speech act as a promise, no acceptance of the terms, and hence no promise.

If this world is near enough to be relevant to the counterfactual connecting MacDonald's maxim to his action, then it falsifies that law. The connection between adopting the fast cash maxim and making the promise is not counterfactually robust enough to be a law, even if it happens to work in the actual world. So we cannot understand MacDonald's action by attributing this maxim to him in an interpretation. Instead we must offer a mechanical explanation, for instance, one that understands MacDonald as overwhelmed by his greed.

This example helps us understand the dynamics of FUL Constitutivism. First, we see ideas about natural law and human nature creep into Kant's view (or my reconstruction of Kant's view anyway). We do not say that one ought to refrain from false promises because it is contrary to laws of human nature, but we will find, as we do in this case, that principles of human nature are the most fertile place to look for facts showing that a given maxim-action connection does not instantiate a law. Thus features of human nature constrain the norms of action indirectly, by constraining which maxims instantiate counterfactually robust laws.

We should also emphasize that the requirement of nearness is not a trivial one. High-level laws, like those that might connect maxims and actions, need not be as counterfactually robust as, say, the laws of physics. Thus for a world to undermine one of these laws in a way that actually handicaps our explanation of a person's behavior,

it must be a very near neighbor in logical space. And looking back at MacDonald's case, it is not clear that the world Kant's analysis point us to is close enough to do the job.

The third point to make about these dynamics involves a familiar problem for the FUL. It seems that MacDonald's behavior fails to be interpretable because it relies on contingent features of the world (people generally keep their promises) that do not hold up in other possible worlds (ones where promise breaking is rampant). But this is clearly the wrong sort of thing to anchor our account of categorical normativity. We need look no further than the all-too-familiar false negatives and false positives that plague the FUL. Had MacDonald included these contingent features as explicit conditions in his maxim—"I will make a false promise in order to get some ready cash, but only if other people accept promises"—then the principle connecting maxim and action would be made trivially true in those worlds where these features do not hold, and so those worlds would not undermine the putative law. On the other hand, some innocuous maxims may fail to instantiate laws just because they are so intimately tied to these contingencies. The maxim, "I will play tennis on Sunday morning in order to avoid the crowds" relies on the fact that most people are not interested in playing tennis on Sunday mornings because they go to church, or sleep in, or are hung over. But if we go to a nearby possible world these Sunday morning contingencies vanish, and the maxim becomes self-defeating.

These problems have the same source. Contrary to Kant's assumption, the maxim is too narrow an instrument of interpretation to carry the burden we have assigned it. A person's maxim may instantiate a law of interpretation even if his behavior in a more general sense cannot be understood as that of an agent, and vice versa, his behavior in general may be interpretable even if the particular maxim we home in on for a particular action does not instantiate a universal law. Kant's focus on maxims does mark an improvement over those theories that take the locus of normative evaluation to be the act itself or the motive behind an act. But not even the maxim's admixture of act and aim is adequate to capturing all the nuances of an agent's psychology.

2.6 Realm of Ends Constitutivism

Everything we did in the previous section was premised on Kant's assumption about the centrality of maxims in our laws of agency. This assumption was helpful because it fixed the possible content of those laws enough to enable us to go from the nebulous Generic Constitutivism to FUL Constitutivism. However, Kant's assumption was also a dubious one. Our practical psychology is too complex, too balkanized, and too multifarious to be shoehorned into a one-concept theory. Once we drop this assumption it becomes difficult to say anything definite about what agency requires; we are back to the thought that one is required to follow the laws of agency—whatever those might be. Still, our hopes of getting a more informative form of constitutivism aren't dashed quite yet.

There may be no principle requiring and forbidding separate lists of actions just in virtue of their connection to agency, but we can get the next best thing. We can derive a rough schema of categorical norms by showing that the demands of agency include a requirement to participate in a 'legislation' of normative principles. In other words, we can show that agency requires us to abide by whatever emerges from some construction procedure. The proposal becomes clearer when compared to other familiar forms of contructivism. Hobbes thinks that we are bound to the outcome of one such a procedure because it is collectively rational. Rawls thinks we are bound to the outcome of another because it enshrines our existing ideas about what is reasonable. By contrast, the thought I am advancing is that we may be obligated to participate in a similar procedure because it is a requirement of being an agent.[29]

The procedure I want to focus on can be found in Kant's discussion of the Realm of Ends. Kant offers a few formulations of the Categorical Imperative; the Formula of the Realm of Ends is one of them. It requires that the principles we act upon survive a particular legislative procedure that we undertake with other agents. In particular, we may only act on those principles that could be enacted as a law for a community of individuals who also are engaged in this activity legislating for an entire community.

[29]This interpretation of Hobbes and Rawls is straightforward. I am less sure about Habermas. See Habermas (1998, pp. 39-45) as well as his 'debate' with Rawls in the same volume.

The FRE does not offer immediate and specific advice to do this and avoid that. Rather, it requires us to participate in a procedure of reciprocal legislation that will eventually result in an exhaustive practical theory, in "a complete determination of all maxims."[30]

In this section I argue that someone is an agent just insofar as she submits herself to this procedure and hews her behavior to its outcome. Earlier I argued that Generic Constitutivism followed from a special self-awareness, namely awareness that the *I* who explains agency and the *I* who act are one and the same. The laws I use to explain agency become normative for me the moment I realize that I am one of the creatures bound by those laws. (This is something like the moment Wile E. Coyote realizes that he's walked off a cliff.) A form of constitutivism for the Formula of the Realm of Ends follows from the further awareness that *everyone else* is engaged in precisely the same endeavor—interpreting others' behavior, systematizing those interpretations into rudimentary laws, acting in accord with the laws so determined—and that my efforts are inextricably bound up with theirs. This entanglement means that success in this endeavor depends on the very same coordination demanded by the FRE.

Let's organize this Mutual Interpretation Process (as I will call it) into steps:

1. Every person acts, and in so doing they determine the phenomena to be explained by a theory of agency: the behavior present in the world.

What is essential about this step is that the target explanandum of our theory of agency, and so the primary constraint on our conception of the laws of agency, is constituted by the actions of people. Thus the phenomena we are trying to explain is partially man-made.

2. Every person observes this behavior, and they all try to articulate an explanation of each instance of behavior by assimilating it to the laws of agency they believe to obtain.

There are practical limitations to how much behavior can be observed, as well as all the usual limitations on our ability to systematize and explain. These affect our

[30]See 4:433-6. Also see Rawls (2000, p. 208) for a fleshing out of this proceduralist interpretation of the Realm of Ends.

interpretations in obvious ways. A woman born in present day Brooklyn is unlikely to have much experience with how the women of ancient India express their grief, and so her theory of agency will necessarily be a parochial one. It may, for instance, deem a practice like *sati* inexplicable only because of a lack of data. But the same limitations exist for all sciences.

Next, we must understand that the laws that we rely on are only hypotheses. They are not fixed throughout all stages of interpretation, and sometimes we will revise our picture of the laws of agency if we can better explain the phenomena.

3. If an individual can get a better explanation of the behavior she observes by revising her theory about the laws of agency, then she will make this revision.

For example, suppose an interpreter has a theory of agency according to which performing a certain action from a certain motive unintelligible. What happens when our interpreter finds what appears to be someone performing this action out of this motive? He can do two things. He can give a different interpretation of this behavior, one on which this person's action is, despite appearances, intelligible, or he can deem this action genuinely uninterpretable and resort to non-interpretative explanation. These two strategies are usually appropriate, but what if our interpreter keeps seeing this kind of action? Our interpreter, like any other scientist must stop explaining away this data as misleading or erroneous and see it as genuinely recalcitrant. And this should lead him to revise his theory of agency. What seemed to be the best explanation for the phenomena he had witnessed now seems inadequate, and so he should revise it. But this means that the behavior of *other people* has an influence over our respective accounts of the laws of agency, for it is this behavior that (partially) constitutes the phenomena those laws aim to explain.

4. Each individual then acts, which means checking her proposed course of action against what she takes the laws of agency to be.

Just as playing chess means checking one's proposed move against what one thinks the rules of chess are, making a move in the game of agency requires checking one's action against the laws of agency. This does not mean that a person's theory of the

laws of agency completely determines what she does; there are questions of chess strategy and life strategy. Our theory of the laws of agency will first and foremost function as a filter. If I am tempted to run screaming from the building because I have caught a glimpse of a spider, but I consult my picture of agency and find that this action would conflict with my agency, then that may—ones hopes—keep my hysterics in check. Admittedly, sometimes this won't work: sometimes I'll run out of the building anyway (if, for instance, I'm overwhelmed by fear).

The metaphor of filtration is too clunky to do justice to the myriad ways that our understanding of other people can inform our own action. I can't give an adequate of all this here, but I'll mention two examples that hint at the complexities:

> Manfred is a junior high school student infatuated with one of her classmates. Sadly for Manfred, her classmate has rebuffed every one of her overtures. Manfred thinks that the only thing to be done is an even more ostentatious demonstration of her affections. But as she waits outside of her beloved's chemistry class, dressed as a sailor and ready to sing her romantic variations on Gilbert and Sullivan's "He is an Englishman", she realizes that her behavior has veered into outlandishness. It's not just that Manfred now finds herself behaving in a way she doesn't like; her love is strong enough to conquer that. She realizes that were she to see this outlandish behavior in another person, she would find it utterly bewildering. She would think that person was overwrought, fixated, around the bend. She would not understand that person as acting for reasons of love but as overcome. So Manfred retreats from her schemes, no longer seeing them as even in the realm of sensibility.
>
> Around the same time Manfred hears her friends talking about how they hate stupid people who nonetheless succeed in their studies. This phenomenon leads Manfred, in the guise of her theory of agency, to recognize the attitude of *resentment*. Such an alteration allows Manfred to make sense of her otherwise unconceptualized pangs of contempt, which, in turn, enables her to resent people. Thus Manfred's understanding of

other people's actions gives her a new way to understand her own, and this gives her new ways to act.

These are just two small examples of our how our provisional theories of agency are used to modulate our own behavior, but I think they give a sense of how this regulation might go.

In doing this everyone jointly determines what behavior exists in the world.

5. Everyone acts, and these actions jointly determine the behavior present in the world.

The process then iterates, and, if we are lucky, tends toward a point of equilibrium, one in which our actions and our theories of agency are all in sync with each other.

...

N. We reach a point of equilibrium when each individual *qua* explainer is no longer obliged to make revisions to her hypothesis about the laws of agency in order to accommodate the actions of each individual *qua* actor.

(It is actually too strong to say that this process must tend toward this equilibrium. I'll say a lot more about this later.)

Rendering the Mutual Interpretation Process into discrete steps is an idealization. In practice, the whole thing is a blur. Everyone is engaged in every step at every moment, the revision of laws of agency is a relatively unconscious process, deliberation does not involve any active comparisons to one's theory of agency, and each person is unaware of the behavior of most other individuals. What the idealization brings out well is a web of interdependencies. Schmidt's actions depend on his understanding of agency, which in turn indirectly depends on Schumacher's actions, which in turn depend on *his* understanding of agency, which in turn depends on the actions of Schmidt, Schuler, Schoperlogen, and a whole motley of other agents. Thus what we see is a kind of *holism* of Schmidt's actions, Schmidt's theory of agency, Schumacher's actions, Schumacher's theory of agency, and so on. Anything affecting one part of this web will spread through these connections until it has affected all the rest. We can

therefore think of these individuals' actions and their theories of agency as constituting a single 'web', à la Quine's web of belief. And when there is agreement amongst this web's many nodes—when we have equilibrium—it constitutes agency.[31]

This relationship makes agency very special. When we do physics, we are trying to explain phenomena that are very much independent of us, phenomena it would be inappropriate to manipulate in search of a different theory. Doing this would, in effect, mean deceiving ourselves about the real nature of a physical world independent of us. We would be hiding something from ourselves that is there anyway. But the case of agency is different. *We* produce the phenomena that are to be explained by the laws of agency. Our actions are the essential constituents of the target phenomena, the thing a theory of agency is about. Thus, quite unlike physics, it is in the nature of our theorizing about agency that its phenomena depend on what we do.

Let me sum up where things stand. We are required to adhere to the laws of agency. But to live up to these laws we have to figure out what they are. And when we submerge the dual processes of acting and interpreting into a social medium, one in which everyone is simultaneously acting and interpreting, the result is the Mutual Interpretation Process. Now I suggest that this Mutual Interpretation Process simply is the Realm of Ends procedure by a different name. We can make out the equivalence in three steps.

First, acting on a principle amounts to an attempt to legislate that principle as a law for the entire community because doing so is *ipso facto* the introduction of corresponding behavioral phenomenon into the world. And so this kind of action exerts an influence over others' actions insofar as their account of agency must grapple with this new phenomenon. This influence is small and indirect, but when accumulated over the many actions individuals perform, it adds up to something significant. Suppose Tex rides a mechanical bull to impress a lady. Doing so amounts to an attempt at legislation, in Kant's sense, because it partially constitutes the phenomena that laws of agency are supposed to be laws *of*. By riding the bull to impress a lady, Tex introduces

[31] "Total science is a field of force," Quine (1951) said, and "a conflict with experience at the periphery occasions a readjustment in the interior of the field." I am urging much the same picture for the relationships between different individuals' actions and their theories of action.

this action into the world of behavior as a sensible thing for a man to do, given his background beliefs and desires. And doing that is tantamount to proposing, in the context of a Kantian legislature, a practical principle connecting those background attitudes to that action.

Second, whether or not a principle is ratified by Kant's procedure can be understood as the question of whether there is uptake of Tex's action, which is ultimately a matter of how other agents respond to his antics. Other interpreters *cum* legislators can reject Tex's principle by adopting theories of agency that do not recognize Tex's behavior as intelligible (of course this is not something entirely at their discretion), and they can hammer home this verdict by acting on principles that are incompatible with the one Tex acted on. The most obvious examples involve principles that undercut the whole idea that bull-riding is a sensible way to woo a lady. Acting on these principles introduces behavior into the world that places demands on our putative laws of agency that are incompatible with the demands Tex's behavior imposes. And so the more of this behavior there is, the greater pressure on our theory of agency there is to understand Tex as an anomaly, as a crazy person whose bull-riding is pathological. Thus it is through the actions that we perform and the ensuing empirical pressure they exert on our theories of agency that practical principles are negotiated in the way that the Formula of the Realm of Ends requires.

Finally, the state of equilibrium that the Mutual Interpretation Process seeks is the same state as the total agreement of practical ends that is the Realm of Ends. The key to this equivalence is noticing that both the equilibrium state that the Mutual Interpretation Process seeks and the Realm of Ends are defined in terms of a single ideal of systematicity. Kant describes the Realm of Ends as a "systematic combination of wills under communal laws."[32] Our theory of agency seeks the very same systematicity and lawfulness, only it seeks them as the demands of explanation. Thus the Realm of Ends and the systematic theory of agency are two ways of describing a single state of systematic equilibrium, one that represents it as a practical ideal and one that represents it as a theoretical ideal.

To restate the whole argument briskly: An individual is obliged to submit her-

[32]4:433.

self to the FRE procedure because the material that the laws of agency are laws of—action—is jointly constituted by all the actors of the world, and so the way a person lives up to the requirement to adhere to the laws of agency is by negotiating with her fellow actors in the way set out by the FRE. When we are trying to explain something that is jointly constituted by us and our peers, that explanation becomes a kind of legislation, precisely the kind found in the FRE. And this gives us FRE Constitutivism: the view that the demands of agency commit us to the schematic normativity of the FRE.[33]

2.7 Local and Global Equilibria

One immediate, and perhaps surprising, consequence of my case for FRE Constitutivism is that the laws of agency depend a great deal on circumstance and the predilections of would-be agents. There are no contentful categorical norms (like injunctions against killing and stealing) laid out in advance of human interaction. These norms are reached as people come to mutual understandings about which kinds of action make sense, and this makes it possible that different groups of agents' negotiations may result in very different sets of categorical norms.

We can bring this out nicely with a pair of examples. First: A man kills an acquaintance in a dark alley so he can steal a tidy sum of money that his chum has just acquired by pawning his watch. Second: The same man lies to his most loyal friend as part of an intricate scheme to free himself from a marriage. Now imagine these two acts submerged into two different worlds. The first is the genteel world of a P. G. Wodehouse novel, and that the two men in our examples are, say, Bertie Wooster and his valet Jeeves. In this world, the first action is utterly inexplicable. The

[33]Let me mention a few antecedents to this kind of view, in case they lend it any clarity. The kind of negotiation that I have in mind resembles John Maynard Keynes's "beauty contest" (from chapter 12 of the *General Theory of Employment, Interest, and Money*), a contest whose goal is to evaluate faces according to the standards of other rankers engaged in the same contest, and thus one whose dynamic tends toward coordination rather than arriving at some external standard. Another, older example of this kind of negotiation is found in Adam Smith's *Theory of the Moral Sentiments*. Smith argues that we have a natural desire to be sympathized with and a corresponding inclination to sympathize, but also not to extend our sympathy to sentiments we find incongruous. In conjunction, Smith argues, these two forces regulate the sentiments we have.

members of this world have legislated, through their actions, reactions, and evaluations that murder for money is a shocking and incomprehensible thing. In this world, Bertie's murdering Jeeves would mean running afoul of a law of agency and therefore the violation of a deep and categorical norm. The second scenario, deception, is something easily understood in this world. Indeed, Bertie regularly lies to get out of engagements. He is aware of the dangers of this lying, so he exercises some caution. He may even say that lying is immoral. But for Bertie these thoughts have all the force of an observation about the requirements of etiquette. So in Wodehouse's world there definitely are principles in the air that allow us to make good sense out of Bertie's lie, which means that he hasn't violated a categorical norm. Now, on the other hand, suppose our examples involve gangsters from the world of Personville (or 'Poisonville' as the locals call it), the shabby Montana town that serves as the setting of Dashiell Hammett's *Red Harvest*. In this town it is common enough to kill a man for pocket change that its residents—even the crooked police and town elders—happily cite this connection to explain people's actions. And they do so without astonishment or even much indictment. In a town like Personville it just *makes sense* to kill a man for money, and while many would refrain from doing this because they try to adhere to some code of right conduct, they can imagine circumstances where, despite their code, they might submit to the irresistible logic of murder. On the other hand, for all its violence and depravity, Personville is bound by primitive concepts of honor, and a lie—especially to one's friend and especially on such of such small consequence—is just as incomprehensible in its world as murder is in Wodehouse's.

These examples suggest how the character and dispositions of the inhabitants of a world could diverge in their development in ways that produced a corresponding divergence in the nature of agency in that world. This feature of FRE Constitutivism strikes me as perfectly appropriate. Just as the evolutionary history of amphibians affects the nature of frogs, the social history of humanity affects the nature of agency. The only difference between the two cases is that the latter history is one shepherded by our capacity for reflection and self-correction.

But some will find this contingency and the threat of normative relativism it entails disturbing. There are both good and bad reasons for this alarm. One bad reason

is an expectation that there is a roster of very specific categorical norms chiseled into the foundations of the universe in a way that makes them independent of the vicissitudes of human nature and interaction. This is an odd thing to expect. Normativity should have the same degree of contingency, *a priority*, and fundamentality as human activity itself. It would be surprising if the laws of morality turned out to be more fundamental than the laws that shaped human nature. We can reasonably hope that the norms of morality and rationality are not held hostage to the mercurial ebb and flow of a single man's passions, but trying to secure for them the same status as particle physics is to reach too far.

But there is also a more reasonable concern about this relativism. We think that the people of Wodehouse and Hammett's worlds can *do better.* And we worry that a view like FRE Constitutivism robs us of this ability because it insists that there is no more to normativity than the consensus that agents negotiate between themselves. But if this is the case, how do we criticize consensuses that are intuitively bad? We see the problem acutely if we think about moral trailblazers. Rosa Parks takes it on herself to move society closer to the Realm of Ends, but her action runs contrary to the prevailing norms of Montgomery, Alabama, so it is hard to make sense of the idea that Parks's disobedience is a worthy thing. Finally, I think we see this same concern on a grander scale animating some of the early writings of critical theory. Here, for instance, is Horkheimer suggesting that a Kantian contractualism can fail if introduced into a social medium that is already dominated by a certain ideology, for instance one of bourgeois capitalism:

> The categorical imperative holds up a "universal natural law", the law of human society, as a standard of comparison to [the bourgeois] natural law of individuals. This would be meaningless if particular interests and the needs of the general public intersected not just haphazardly but of necessity. That this does not occur, however, is the inadequacy of the bourgeois economic form: there exists no rational connection between the free competition of individuals as what mediates and the existence of the entire society as what is mediated. The process takes place not under the control of a conscious will but as a natural occurrence. The life of

> the general public arises blindly, accidentally, and defectively out of the chaotic activity of individuals, industries, and states. This irrationality expresses itself in the suffering of the majority of human beings.[34]

I have more to say about how Horkheimer sets about resolving this tension, but the for the moment I think his disagnosis fits in with our general problem. We want to hold onto the critical perspective that allows us to say that there is something wrong with Wodehouse and Hammett's worlds, and with bourgeois Europe, but something right with Rosa Parks's disobedience.

The way to do this is to understand that despite the appearance of an equilibrium, there is something *unsettled* about the constitution of agency in these worlds. There is enough consensus about how people should behave that we can recognize genuine regularities of action and use them for interpreting behavior. But I submit that this consensus is only partial, and so the corresponding legislation of norms is unstable. Indeed the instability lurking beneath the surface of these kind of societies is a major theme of hard-boiled fiction like *Red Harvest*. Hammett, a former Pinkerton detective, thinks that unchecked vice, even if it manages enough stability to resemble a civilization, is a corrosive force and ultimately incompatible with understanding ourselves as human beings. "This damned burg's getting me," mutters Hammett's protagonist, "if I don't get away soon I'll be going blood-simple like the natives." Hammett's solution to the problem of Poisonville is straight out of Hobbes: the restoration of order in the form of a single, ruthless town elder.

I propose we understand this unsettledness in the following terms. I described the process of interpretation as being driven toward an equilibrium point at which action and theories of action coincide with each other. Wodehouse and Hammett's worlds, as well as Montgomery circa 1950, represent what we might call *local equilibria* for this procedure. There is enough agreement between the actions and theories of agency held by the various members of these worlds that any deviation from this rough agreement amounts to a departure from agency and a violation of a categorical norm. This negative reinforcement mechanism cultivates homeostasis. But this state is still

[34]Horkheimer (1933, p. 19-20).

only a local equilibrium because there are states of even greater agreement between individuals' actions and their theories of agency. These states are more systematically organized and organically unified than those in Wodehouse and Hammett's worlds. And in this sense they are more stable. The greatest of all these greater equilibria, what we might call a *global* equilibrium state, is the Realm of Ends itself. There is an unfortunate tendency to understand the Realm of Ends as the "liberal's idea of a good society" (in R. M. Hare's words). In fact the ideal is much more demanding than that. As Allen Wood puts it, "the idea of a "realm" requires a harmony or even an organic unification of ends so that the ends of all can be pursued in common. Even the most liberal society would still be far from achieving the total unity and unanimity required for a realm of ends." Thus we would not find the kind of dissembling, perfidy, and injustice in the Realm of Ends that we find in these local equilibria.[35]

A world mired in a local equilibrium is like a stick pinned against the edge of a river by the back current—a negative reinforcement mechanism—it creates itself. The stick would be in a more 'stable' (low energy) position were it to continue along with the flow of the river and reach the open ocean. But every nearby point in the river is *less* stable than the local equilibrium created by the stick's own back current. And so it remains. So the question we have to answer is, where does our putative obligation to move from a local equilibrium to the global equilibrium of the Realm of Ends come from? If all normativity is generated by the laws of agency that emerge from an equilibrium state, then what is the source of the push from local to global?

The question of why we should move beyond a local equilibrium and strive for the ideal of a Realm of Ends is wrapped up with a familiar question in the philosophy of science. Why should we move from an adequate, but not-so-systematic theory to a more systematic and explanatorily unified one? In the first *Critique*, Kant says that it is reason that pushes us toward systematicity.

[35]Wood (1999, p. 166). I think Kant has an idea of biological harmony (a forerunner to the notion of homeostasis) in mind here. He imagines the Realm of Ends embodying an ideal on which humanity functions as a unified organism, not just a collection of people who manage to get along. This idea gives us a social conception of teleology. The demands of morality are not, as Aristotelians think, connected to the demands of self-preservation for individual human organisms, but to the self-preservation for all humanity understood as an organism.

> If we survey the cognitions of our understanding in their entire range, then we find that what reason quite uniquely prescribes and seeks to bring about concerning it is the systematic in cognition.
>
> The law of reason to seek unity is necessary, since without it we would have no reason, and without that, no coherent use of the understanding, and, lacking that, no sufficient mark of empirical truth; thus in regard to the latter we simply have to presuppose the systematic unity of nature as objectively valid and necessary.[36]

According to Kant, the entire business of understanding is driven by the demands of reason, and reason aims for systematicity. Without reason we could give no explanations, formulate no necessary laws, and achieve no understanding. For these activities come not from our passive sensory powers, but from the reflective power of reason that puts questions of explanation to us in the first place. Thus we are obliged to strive for systematicity beyond local equilibria and the theories of agency that accompany them because reason itself aims for systematic organization. Modal notions may help illuminate this idea.[37] Our stick pinned to the riverbank is in an equilibrium state, but some interventions will break this equilibrium, for instance if a bear comes along in search of fish and knocks the stick away. But the equilibrium achieved when the stick reaches the ocean is stable under these kinds of interventions. Similarly, Hammett, Wodehouse, and Parks's worlds have a measure of stability, but they are not stable under certain interventions. Indeed, Parks and Hammett's Continental Op are themselves responsible for interventions that knock their respective worlds out of their precarious partial equilibria. So the demand for systematicity that reason imposes can be understood, at a first approximation, as a demand for the kind of robustness in the face of outside interventions that these theories lack.

But isn't it illegitimate to solve our problem in this way, to say that the demands of theoretical reason are what drive us toward the Realm of Ends? For one thing this

[36]A645/ B673 and A651/B679.

[37]Compare this thought to the discussion of counterfactual robustness for FUL Constitutivism.

has the counter-intuitive consequence that theoretical reason drives practical reason. It is plainly wrong to understand Rosa Parks's action as arising from a desire for a better theory of agency than the one appropriate to segregated Montgomery. Second, this suggestion seems to violate the spirit of constitutivism by introducing reason's demand for systematicity as an additional normative requirement, one apart from the requirements arising from the constitution of agency.[38]

These would be fatal objections were it not for our special relationship to the phenomenon of agency: in theorizing about agency we are studying something that we create through our actions. Recall that what we originally used to distinguish agency from other aspects of human nature—from our tendency to commit the conjunction fallacy—is our ability to regulate our behavior through the use of reason. We are not slaves to our instincts. We can stand back from our inclinations and hold them up to scrutiny. Our ability to do this is a power of reflective reason. This capacity is not the be-all, end-all of agency, but it is the *sine qua non* of agency, and it is the thing we use to latch onto it. Reason is active both in producing the phenomena of agency and in our theorizing about these phenomena. So when we are offering an explanation of agency that is guided by reason's demand for systematicity, we are not "projecting order" on a nature that otherwise lacks it (as we might be tempted to characterize reason's imposition of systematicity over other parts of nature). We are dealing with the part of nature that this power of reason is responsible for.

Nor are we introducing a new source of normativity when we talk about the regulative ideal of systematicity. Because reason-guided action is our explanandum, the demands of reason must be validated by all the various equilibria that the FRE procedure might come upon. These demands are the one kind of norm that will be common to absolutely all possible systems of norms arising out of equilibria because all such equilibria must, by definition, be under the sway of the critical power of reason. (The demands of reason are, to adopt the nomenclature of supervaluationism, 'super-norms'.) So the requirement to move beyond merely local equilibria is already

[38]The first of these is a common objection to Velleman's theory of practical reason. Thanks to Michael Bratman asking a question that made it clear to me how my account could be accused of the same problem. See footnote 23 above for a longer discussion of the differences between my views and Velleman's.

implicit in the norms ratified by those equilibria themselves.

For the same reasons, our reliance on the regulative ideal of systematicity is not an illegitimate invocation of theoretical reason, but a perfectly appropriate application of reason in a form that is at once theoretical and practical. There are not two faculties of reason, each with distinct demands and kinds of normativity. There is a single faculty of reason that may be applied to different subject matters. And agency is one subject matter where the theoretical and practical use of reason come together. For here we find the reciprocity I have emphasized again and again: our deliberations about what to do affect our inquiries into the world, and these inquiries directly affect our deliberations about what to do. (We could say here that this is where distinctions about direction of fit collapse: our interaction with our own agency is both mind-to-world and world-to-mind, both creation and tracking.) So the systematicity that reason demands and we find enshrined in the Realm of Ends is not a distinctively theoretical or practical ideal, but an ideal of a single, unified reason.[39]

From this picture we get a two-tiered system of norms. Reason gives us a regulative ideal of systematicty that is both theoretical and practical. This ideal acts as a kind of higher-order norm that regulates the Mutual Interpretation Process. This process then produces a constitution of agency, which engenders all our first-order norms.

I mentioned before that critical theory, at least in Horkheimer's early works, was concerned with something like what I have characterized as local equilibria. Now I want to claim another parallel. The solution that Horkheimer proposes to this problem is a program of social research aimed at explicating the dysfunctional order of society. Throwing light on this wayward social structure will in turn 'emancipate' people from that the ideological grip of that structure.[40] Thus Horkheimer, as I understand him, has in mind a research program with an essential practical aim. He is not merely proposing something, like research into the human genome, that is theoretical but may have some instrumental use later on. Nor is he suggesting that it is impossible to keep pragmatic concerns from encroaching on the pristine concerns

[39]I argue for this unity of reason at much greater length in chapter 1.

[40]Horkheimer (1937).

of science. Rather, the *essence* of his program is the practical goal of emancipation. I think for this practical conception of a research program to be legitimate, we must adopt a picture of the reunion of theoretical and practical reason like the one I am proposing.

This solution to the problem of local equilibria also gives us a way to understand moral trailblazers. When Rosa Parks or Gandhi acts contrary to the established norms of one of these local equilibria, we should understand their actions as specially self-referential acts. They amount to a sort of performance art that exemplifies the gap between the way society is actually structured and the ideal of the Realm of Ends. Thus these sorts of actions should not be interpreted with the mundane norms appropriate for riding a bus in Montgomery or acquiring salt in India, but with the norms we use to make sense of artists advancing our understanding. The awareness of these shortcomings then pave the way for collective action that can effect the large scale shifts in behavior that we need to lift a society out of a local equilibrium.[41]

It is worth emphasizing that for all I have said the Realm of Ends is still just a regulative ideal. It is not merely something that we will probably never reach; reaching it is an impossibility for the very principled reason that we are flesh and bone creatures immersed in a messy world. For the same reasons I doubt (*pace* Lewis) that we can ever eliminate all indeterminacy in a theory of agency or all traces of the relativism of path dependence. The phenomena simply aren't up to it. But, again, I think the familiar roster of natural sciences share the same fate. We will never achieve perfect systematicity in physics or chemistry. The phenomena aren't up to it there either. And arguments like Putnam's model theoretic argument and Goodman's grue paradox show that indeterminacy and path dependence, at least when it comes to our choice of concepts, are ineliminable features of inquiry. Many readers will disagree about the upshot of these arguments, but my point is just that agency and the norms its grounds will share a common fate with the natural sciences. In both cases harmony and systematicity operate as regulative ideals even if those ideals are quite impossible

[41]This analysis is probably appropriate for Lady Gaga and other inscrutables. The account of the arts and their role in advancing the understanding I have in mind is roughly that of Nelson Goodman (1976). For an account of philosophy as performance art see the fascinating discussion of the Cynics by Ineke Sluiter (2005).

to achieve.

2.8 Conclusion

I have sought to develop a version of constitutivism that does not rely on an other-worldly conception of agency, but still produces contentful norms. In doing so I have reached a view that has aspects of both realism and constructivism. The laws of agency are genuine laws of nature, but they are laws of nature partially constituted by human beings. These laws have a moral character (in the sense that they require us to care about the needs of others) because agency is constituted by a of negotiation amongst different actors, a process quite like the legislation we find in Kant's Formula of the Realm of Ends.

Chapter 3

Transcendence

We think morality has two crucial features, but these features are hard to put together. We think that morality must be humane and that it must be worldly. By 'humane', I mean that we want morality to really be *ours*: to be grounded in human attitudes and activities, to be something within our practical and intellectual reach, to make some real difference to human beings. By 'worldly', I mean that we want morality to be *more than merely ours*: to be something greater than the prejudiced bickering of our parochial opinions, more than a self-satisfied coherence project, more than a chase after our own tail. We want morality to give us guidance in becoming *better than* we already are. So, on the one hand we think it inconceivable that a heavy isotope of Berkelium might turn out to be the one thing with intrinsic value. And on the other we think that value is not just the shadow of desire, but something that desire aspires to.

We hold out the same hope for the main concepts of epistemology—justification and knowledge. We resist the coherentist picture of justification on the grounds that it is not worldly enough. Someone could have a completely consistent picture of the world but be so out of step with reality that he is utterly unjustified in holding it. And yet, at the same time, we think that justification must regulate inquiry, and we wonder how the nature of things beyond our ken could possibly do that.

Morality and inquiry are both normative domains, and so it seems reasonable that the demands of worldliness and humanity are ones common to all normative notions: value, justice, right, justification, validity, reasons—just to name a few. In this paper

I talk collectively about these normative items and the concepts that pick them out. To do that I commandeer the word *Goods*.

The problem I take up in this paper is that these two demands on our Goods seem damned impossible to satisfy simultaneously. We can be old-fashioned sentimentalists about morality and get a thoroughly humane theory, but only at the risk of tying morality too closely to the contingency of our passions and sacrificing its worldliness. Or we can be moral realists and make morality just as worldly as physics, but then we tempt questions about what this arcane system of duties has to do with actual human beings trying to live their lives.

This conflict goes back to a question Socrates put to Euthyphro. Socrates asks whether the gods value piety because it is good or if piety is good because the gods value it. Socrates's question naturally generalizes. Instead of asking about value, we can talk about any Good. Instead of asking only about valuing, we can ask about any attitude or activity. Instead of talking just about the gods, we can talk about any class of individuals—trained value arbiters, Bostonians, human beings, rational creatures. If we make all these generalizations we get the following pair of questions:

> Does something possess a Good (e.g. is it valuable) because of the activities and attitudes of some class of individuals (e.g. their valuing it)? That is, is it *Socrates Subjective*?
>
> Or does it possess a Good independently of the activities and attitudes of any class of individuals? That is, is it *Socrates Objective*?[1]

If we answer 'yes' to the first question, then we do well by the demands of humanity, but we also abandon any pretense of worldliness. If, on the other hand, we answer 'yes' to the second question, then we get the reverse.

[1]How to analyze the 'because' relation here has been an issue of controversy lately. Some writers (see, e.g., Crispin Wright (1992a)) try to cash it out in epistemic terms, as a certain relationship between individuals and Goods holding *a priori*. Others try to cash it out in terms of asymmetric supervenience relationships. More recently there has been a resurgence in metaphysical approaches. These posit a primitive 'grounding' relation between the relata of 'because' (see Gideon Rosen (2010)). I don't find any of these approaches particularly promising, so I lay this debate aside and leave 'because' blissfully unanalyzed.

Here is the deep problem. If Socrates's pair of questions is a proper dilemma, and if each horn of the dilemma delivers one desideratum while denying us the other, then we're in trouble. We will always find ourselves estranged from either worldliness or humanity, and that fate is intolerable.

This tension shows up in many places. In epistemology it appears as a duel between two sets of intuitions, the externalist thought that the goal of inquiry is to be well-connected to the world—by having knowledge, for instance—and the internalist thought that inquiry has to provide guidance by setting out rules. In moral philosophy we see the tension in what Michael Smith calls ' the moral problem'. We want to preserve the motivational features of morality without sacrificing its categoricity, but this is hard to do.

A lot of work has been done trying to decouple Socrates's two choices and our two intuitive desiderata. Smith's own work is a good example of this. He seems to believe that morality is Socrates Objective, but, as he painstakingly argues, because certain statements about our relationship to these objective moral facts are analytic, we are automatically motivated by our recognition of them.[2] On the other hand, many Socrates Subjectivists argue that the subjective features of human beings that ground morality are general enough to make morality worldly.

Despite the prominence of this sort of project, no one, as far as I know, has taken on the more ambitious task of denying Socrates's Dilemma itself. No one has tried to show that there is a middle ground between Socrates's choices.[3] My goal in this paper is to do that.

This may sound impossible, as it certainly appears that the natural reading of Socrates's choice partitions logical space. Either something possesses a Good *because*

[2]Smith (1994, ch. 6).

[3]Any claim this sweeping requires some caveats. Some writers, notably theological voluntarists, have claimed that they side-step Socrates's Dilemma by affirming both horns of it. This is a view we find, for instance, in Duns Scotus and Crusius. (Maybe in Pufendorf too, but I'm less sure.) These writers claim to avoid the anti-voluntarist implications of Socrates's Dilemma by making God's will conform to the good, which allows them to say that things are good both because God wills them and in themselves. But this is not a real evasion of the Socrates Dilemma. It just opts for Socrates Objectivity and inserts God as a middle-man between moral agents and the good. This may be dialectically useful for religious apologists, but it does not dispatch with the Dilemma itself.

of us or *independently* of us. My strategy in this paper is to show that the way we understand "us" in this context is a more subtle issue than we realize, and that this subtlety gives us the latitude we need to find a middle road between Socrates's choices, one which offers a vision of our Goods that is at once humane and worldly.

3.1 Constructivism and the problem of transcendence

I want to begin by looking at the triumphs and trials of one program that has tried to join worldliness and humanity: constructivism. Traditional constructivists believe that a theory of a Good can be Socrates Subjective while still being worldly. The strain of constructivism I want to use as a stalking horse is John Rawls's political constructivism. Rawls says that "the concepts of political justice (content) may be represented as the outcome of a procedure of construction."[4] Importantly, this procedure is not a crude process of consensus-building. Instead it is designed to "model the principles of practical reason in union with a conception of society and person"—particularly through the idealization offered by the Original Position—to arrive at a set of principles that could be reasonably adopted by a society.[5]

Nonetheless, Rawls is quite ready to affirm Socrates Subjectivism. This comes out when Rawls describes the difference between political constructivism and its foil, rational intuitionism:

> The intuitionist regards a procedure as correct because following it correctly usually gives the correct independently given judgment, whereas the political constructivist regards a judgment as correct because it issues from the reasonable and rational procedure of construction when cor-

[4]The final version of this theory, as far as I know, is Lecture III of Rawls (1996a). There are many examples of constructivism in other domains. For constructivism about value see Korsgaard (1996b); about the moral law see Rawls's sixth lecture on Kant in Rawls (2000); about moral matters in general see Korsgaard (2003); about validity see chapter three of Goodman (1955); about epistemic justification see chapter four of Elgin (1996); about reasons see Street (2008). There are of course many other views that have an outlook similar to constructivism, and thus face companion objections, but pursue the details in different ways—salient examples include ambitious forms of non-cognitivism and analytic functionalism.

[5]Rawls (1996a, p. 91).

rectly formulated and correctly followed.[6]

Thus, if a procedure accurately models practical reason, and is founded on the right notions of person and society, then Rawls is committed to Socrates Subjectivism about that procedure. Something is just *because* it is so called by this procedure.

I take this to be a defining characteristic of all constructivisms. The constructivist's procedure may be sophisticated—and Rawls's is certainly that—and it may involve iterated stages of idealization, purification, and equilibration. But once we have this procedure fully in hand and agree on its bona fides, the constructivist must say that for some x to possess good G *simply is* for x to emerge from an appropriate G construction. There is no further question about whether the procedure conforms to a separate standard, and no "independent order" to serve as a model. And this means that all constructivisms will land on the subjective side of Socrates's Dilemma.

And yet for all that, a well-wrought constructivism like Rawls's seems to go a long way toward giving us a picture of Goods that is at once humane and worldly. Rawls's conception of justice is clearly humane, grounded as it is in a rich conception of human beings as practical reasoners with their own values immersed in a social medium. But we can also make a case that it is worldly. First of all, Rawls's reflective equilibrium procedure encompasses the input of umpteen different people (and Kant's putative constructivism goes so far as to include all 'rational beings'). Second, the fact that Rawls's construction involves an idealization—the Original Position—ensures that this procedure is not just a way of regurgitating personal opinions. Thus the procedure issues in a broad-based authority that trumps any individuals' opinions about justice.[7]

So Rawls's conception of justice is Socrates Subjective, but far from vulgarly parochial. This invites the question of the moment: does the unparochialism of Rawls's constructivism make it *worldly* in the way we want? We can ask much the same question in terms of objectivity: Rawls's theory achieves one grade of objectivity. Rational intuitionism offers another, seemingly stronger grade. Which of these

[6]Rawls (1996a, p. 96).

[7]cf. the discussion on Rawls (1971/1996b, pp.516-517).

grades do we need? Is Rawls's good enough, or are we inevitably pushed toward Socrates Objectivity?

Which grade of objectivity we need will depend on what we expect objectivity to do for us. Rawls himself outlines five ways that we might need the principles of justice to be objective, and he shows how political constructivism succeeds on each count.[8] If someone wants to show that Rawls has failed she must produce some useful conception of objectivity that his constructivism isn't up to. She has to point to an itch that constructivism cannot scratch. Many objections to constructivism never get off the ground because they don't play this game.

But there is one sterling exception to this pattern. It is part of our very idea of a Good that it *transcends* us, our activities, and our attitudes. Thus, the objection goes, our most basic acquaintance with concepts like justice, value, and justification involves the thought that our inquiry into these Goods aims at something beyond our own attitudes and activities, beyond anything that can be captured by a construction procedure. So no matter how broad-based, unparochial, and sophisticated we make our deliberations about whether something is just, valuable, or justified, there will always be a further question: our methods call this thing valuable, but is it valuable? Critics of constructivism offer the fact that this kind of question is always open as a self-evident datum about our normative thought. At no point in our refinement of our construction procedure do we stop asking whether we have really gotten onto value, and this makes our inquiries into Goods *open-ended*. In the course of these inquiries, we may get the right answer about value, and we may get it with a great deal of certainty. But even if we reach this happy state, these inquiries will go on, because there is a gap between our practice of valuing and the things that 'ground' facts about value. And this gap is one into which we can introduce the doubt that keeps inquiry churning.

There are numerous examples of this kind of argument. The best known may be a primitive rendering of Moore's Open Question argument. Our practice of valuing may say that this thing is valuable, and our construction procedure for justice may say that this social order is just, but, as Moore argues, there is always the further

[8]Rawls (1996a, pp. 110ff).

question, *Are these things really just and valuable?* And because this question always has intuitive force, no matter how broad our construction procedure, constructivism can never succeed. (We can accept that the Open Question is always live without buying into the dubious particulars of Moore's argument.) We see a similar thought from David Wiggins, one wielded against non-cognitivism, but which could just as easily be directed at the constructivist.

> By the non-cognitivist's lights it must appear that whatever the will chooses to treat as a good reason to engage itself is, for the will, a good reason. But the *will itself*, taking the inner view, picks and chooses, deliberates, weighs, and tests its own concerns. It craves objective reasons; and often it could not go forward unless it thought it had them.[9]

R. Jay Wallace seems to think that transcendence is one way we separate facts about the dispositions of our psychology and political institutions from norms about what what *ought* to be. By denying transcendence, constructivism imperils its status as a theory of normativity.

> In its attempt to anchor normativity in the agent's actual volitional commitments, constructivism in its purest form ends up collapsing the important distinction between norm and psychological fact, ought and is, in ways that call into question its credentials as a theory of normativity.[10]

Finally, Allen Wood makes the same point very well when he emphasizes the open-endedness of moral inquiry in arguing against a Rawlsian interpretation of Kant.

> Kant is a moral realist because realism is the only way of preserving the *critical* stance necessary to all moral thinking, the *open-endedness* of moral inquiry.[11]

All these objections make two points. The first is that it is a basic part of normative thought that the subject matter of that thought is transcendent of it. Thus the Open

[9]Wiggins (1976, p. 99) (my emphasis).
[10]Wallace (MS).
[11]Wood (1999, p. 157).

Question will always have purchase on us, and our inquiries into value, justice, justification, and validity will be open-ended. The second thought is that we have to understand this transcendence in the terms offered by Socrates Objectivity. There is some definite, real nature to value, justification, justice, and validity, but it is always beyond our grasp, so we can never be quite sure whether we are onto it.

I accept the first of these points, the bare observation about transcendence, but not the second. It is tempting to slide from the transcendence of a Good to the thought that there is some independent, and therefore transcendent, nature of that Good. But as a matter of logic, the inference does not go through. Of course, the inference may still go through if there is no way to give an account of transcendence within the scope of a theory eschewing Socrates Objectivity. My task now is to show that such a thing is possible.

3.2 Push and pull

I begin my account with a metaphor. Someone who signs on to Socrates Objectivism thinks that our Goods are transcendent because they are grounded in a transcendent reality that our inquiries are trying to latch onto. They think the open-endedness of inquiry is the effect of our inquiries being forever *pulled* toward this transcendent reality. But wouldn't we see the same phenomenon if we understood our construction procedure as being forever *pushed* from behind (or, stretching the metaphor, from within)? Our Goods would then be transcendent not because they are rooted in something forever obscured from view, but because our reckoning with them is animated by an inextinguishable transcendental *urge*.

An analogy might clarify this metaphor. Aristotle distinguishes between two ways of thinking about infinity. We can think of something as being *actually* infinite if we understand it as a single totality of infinite size. So just as $\{1, 2\}$ has two elements and $\{1, 2, 3\}$ has three elements, we can think of the set of all natural numbers as a totality with infinitely many elements. This is to think of the natural numbers as an actual infinity. On the other hand we can think of a set or series as merely *potentially* infinite. The natural numbers are not a totality with infinitely many elements, but

a series without an endpoint. We can call the actual infinity a positive conception of infinity, since calling something actually infinite is a way of ascribing to it a definite property, the property of having infinity elements. Likewise, we can call the potential infinity a negative conception of infinity, since calling something potentially infinite is just a way of denying something about it, namely that it has an endpoint.[12]

The analogy I want to pursue is this. The old-fashioned realist thinks of the transcendence of our Goods and the open-endedness of inquiry on the model of Aristotle's actual infinity. Just as we can think of an infinite series tending toward an infinite totality, we can think of our inquiries tending toward some transcendent reality. My proposal, by contrast, takes Aristotle's potential infinity as a model: there's nothing that we tend toward in these inquiries, but they are open-ended because they have no stopping point.

3.3 Temporally open-ended constructivism

One obvious way to cash out this analogy is to introduce a temporal parameter to our construction procedure. At t_1 we perform a construction like the one Rawls describes. Then we do it again at time t_2, and again at t_3. Of course, things change between t_1, t_2, and t_3: there are new people with new considered judgments, new social orders we never considered, new systems of value, new ways of conceptualizing relationships. This means we have good reason to expect different outcomes of our construction procedure at different moments. Next we bring these time-indexed constructions into reflective equilibrium with each other. We do this with a higher-order construction procedure, one that takes our time-indexed constructions as inputs. Because time never stops, this higher-order construction doesn't either. Even if we have gotten the same outcome for t_1 through t_{10000}, we must still be open to seeing something at t_{10001} that leads to a revision.[13]

If we construe our construction procedure in these terms, then open-endedness is achieved temporally: for all times t our construction procedure goes on beyond t.

[12]Aristotle sketches this distinction in *Metaphysics*, Θ:6 and again in *Physics*, iii.

[13]cf. Rawls (1996a, p. 97): "the struggle for reflective equilibrium continues indefinitely."

And this approach, it appears, allows us to answer 'no' to both of the questions posed in Socrates's Dilemma. For no time t is it the case that our procedure at time t makes it the case that a thing x possesses a Good G, and thus at no time is our procedure Socrates Subjective. But we are still constructivists, and so we flatly deny that our Goods have some nature in themselves that this procedure chases after. Thus we are not Socrates Objectivists, either. We live someplace in between.

3.4 Socrates's Dilemma restored

There is a problem with this approach. While Aristotle thought there were no actual infinities, conventional wisdom nowadays is rather the opposite. Scarcely anyone thinks there is any problem saying that there are $\aleph_0$ natural numbers and $2^{\aleph_0}$ real numbers in just the way we say there are twelve Apostles and five Marx Brothers. This matters for temporally open-ended constructivism because it means that we can think of our construction procedure as a single totality, not an open-ended thing without boundaries. The clever thought that temporal open-endedness offered the constructivist was the ability to say that at no time did a construction procedure constitute facts about Goods. This enabled him to deny that he was a Socrates Subjectivist. To this a critic will retort that temporally open-ended constructivism is still very much within the spirit of Socrates Subjectivism. And we can show this, she will say, by thinking about *the set of all times* that our higher-order construction procedure ranges over. If we look at our construction procedure as an infinite totality of moments, and we gather all those moments together and think about them as a single class, then we *do* get a 'yes' answer to Socrates's Subjectivist question. Something possesses a Good just because this construction procedure says it does. And this is true even though there is no particular moment when our procedure spits out its answer.

The crux of the problem with temporally open-ended constructivism is that the construction procedure it embraces is large, but still *bounded* (in the way that the set of natural numbers, understood as an actual infinity, is large but bounded). This means that we can still talk about the procedure as a single thing that our Good are subjective relative *to*. This is a decisive blow against temporally open-ended constructivism, but

it offers us a hint for how we might improve on it. We need to find a way to make our construction procedure not merely large, but truly unbounded. The keystone of my critic's case against temporally open-ended constructivism was his observation that even though our construction procedure is very large, we can still understand that procedure as a unified whole. We can still characterize it in a way that allows us to say, "*this* thing, *this* construction procedure is the thing that satisfies the criteria for Socrates Subjectivity."

If we can construe our construction procedure in a way that forestalls this kind of demonstration, then perhaps our construction procedure can avoid Socrates Subjectivity. But this would seem to require us to deny that there is anything that our constructivism makes our Goods relative *to*. That is, it would require denying this principle:

> *All-in-one principle.* It is always possible to think of the construction procedure associated with a Good as a single totality.

The upshot of the All-in-one principle is this. Even if we insist that our construction procedure is infinite in many dimensions (time, space, persons), we can still think of it as a unified totality, and this means that our constructivism ends up committed to the Socrates Subjectivity of our Good. But if we deny the All-in-one principle, then we may can find a third way between Socrates Subjectivism and Socrates Objectivism.

Such a feat requires a foray into the philosophy of mathematics.

3.5 Indefinite extensibility

In 1899 Georg Cantor claimed to have discovered an antinomy in the very idea of certain totalities. "A multiplicity can be such," Cantor wrote, "that the assumption that all of its elements 'are together' leads to a contradiction, so that it is impossible to conceive of the multiplicity as a unity, as 'one finished thing'." Less than a decade later Russell gave an informal characterization of how to produce these antinomies. "There are some properties such that, given any class of terms all having such a property, we can always define a new term also having the property in question." And

this, Russell said, implied that "we can never collect all of the terms having the said property into a whole; because, whenever we hope we have them all, the collection which we have immediately proceeds to generate a new term also having the said property." Some years after that Michael Dummett baptized the concepts producing this arrangement. An "*indefinitely extensible* concept," according to Dummett, "is one such that, if we can form a definite conception of a totality all of whose members fall under the concept, then we can, by reference to that totality, characterize a larger totality all of whose members fall under it." Thus it is part of the nature of indefinitely extensible concepts that their extensions cannot be combined into a single totality. Importantly, this is not the same kind of 'unboundedness' that some writers associate with vagueness. The concept *bald man* is vague because it is difficult, perhaps impossible, to find a set of *all and only* the bald men. But it is not indefinitely extensible because we can easily find a set containing all the bald men: the set of all human beings will do nicely. What is distinctive about indefinitely extensible concepts, if I may be permitted a Cole Porter reference, is that they cry out, with all the fury of semantic antinomy, "don't fence me in!"[14]

Dummett thinks that the concept *set* is indefinitely extensible. The intuitive case for this is easy. If concepts had mottos, then *set*'s motto would be *e pluribus unum*: a set collects disparate things together into a single totality. And this sort of process is not, in principle, bounded. Suppose we take a mouse and Barbarossa and collect them into a set. This set is just as much a thing as its elements are, so it is just as fit to be collected with other things and formed into a new set. And this new set can, in turn, be collected together with other things to form yet another set. The most important instance of this process of generation is the powerset operation: we can 'generate' a new set from an old without the aid of mice or Barbarossa if we take the subsets of a given set and collect those together into a new set. The general procedure here, *generation by collection* we might call it, seems to go on without bound: any set we think is the last or largest can be joined with other things to create an even larger set. Now, one could say that there really is a largest set, a set V of all sets, and all this

[14]See Cantor's letter to Dedekind, which is reprinted in Jean van Heijenoort's anthology *From Frege to Gödel*; Russell (1906); and "What is mathematics about?" in Dummett (1993).

joining together takes place within its ambit. But saying this requires us to say that sets formed out of V and other items (e.g. the set $\{V, Barbarossa\}$ or the powerset of V) will be *inside of* V. This strikes me as a desperate move, one that runs contrary to the concept of set as something *new* that we construct by bringing disparate things together.

We can sharpen this intuition, as Dummett and others do, with Russell's paradox. Form a set r that contains all sets that are not members of themselves, $r = \{x \in V : x \notin x\}$. Since V contains *all sets*, it must contain r. But now we might wonder whether r is itself a self-membered set. By definition, $r \in r$ just in case $r \notin r$. But that's a contradiction, so something has gone awry. The natural suggestion is to reject our strange assumption that there can be a set V containing all the sets. If we take this suggestion, then we regard Russell's paradox as a *reductio* of the hypothesis that there is a totality of all sets.[15]

3.6 Perspectives and their extension

My way around the All-in-one principle involves a construction procedure whose inputs—the things we feed into the procedure—cannot be collected into a single totality because the concept those inputs fall under is indefinitely extensible. This makes such a construction procedure open-ended in a deep way: we cannot even conceive of the completion of such a procedure because it is contrary to its nature that it should be completed.

One way in which constructivism is humane is that its inputs are not propositions floating in the void, but judgments made by people. But most constructivisms also employ idealizations designed to make the procedure something more than mob-rule. Such idealizations require imagining judgments made by people who do not actually exist: no one actually lives behind the veil of ignorance, and so no one actu-

[15]We could instead say that V does contain all sets but deny the legitimacy of forming sets using the comprehension schema used to form r. But this seems unprincipled: restricted comprehension is not ordinarily problematic, and there is nothing intrinsically paradoxical about a set like r. Also, we could continue to insist that all sets can be formed into a single totality while denying that that totality is a set: we can say that it is a proper class. But this kind of book-keeping maneuver only delays the contradiction. We can just as well run Russell's argument for any 'set-like' entity.

ally makes judgments about the social contract from that position. This fact does not, however, amount to a depersonalization of the judgment. These judgments are still personal, even if not made by 'real' people. The way to talk about this is to speak of a judgment being made *from a perspective*. No one actually occupies the Original Position, but we can speak of the judgments made from that perspective. More generally, when no one occupies a given perspective, we can still include it by imaginatively projecting ourselves into it. It is natural, then, to think of the inputs to a construction procedure as considered judgments made from various *perspectives*.

What is a perspective? In a moment I am going to present some difficulties for giving a complete answer to this question. But we can still look at paradigms and gesture at how they might be generalized. Begin with the idea of a spatial perspective. It is something like a vector rooted at a point, pointing in one direction, with a magnitude that indicates its focus. Thus spatial perspectives are distinguished by three parameters: root, direction, and magnitude. Now think of all the other things that can distinguish the not-just-spatial perspective of a participant in a construction procedure: time, background beliefs and desires, ambitions, predilections, biases, cognitive capacities, commitments of conscience, religious affiliations, cultural milieu, ideas of justice, sensitivities of sex, race, and class, and on and on. Each of these parameters can take on several different values, and a perspective is a tuple of these values.

I can now state my way around the All-in-one principle. First, the concept *perspective* is indefinitely extensible in the same way that *set* is. Second, we can imagine a construction procedure that takes all perspectives as inputs. (It may discount some of them rather heavily without ever outright excluding them; I'll come to this point later.) These two facts allow us deny the All-in-one principle—the claim that it is always possible to form our construction procedure into a single unit—and this, in turn, gives us the ingredients for a truly open-ended constructivism.

Perspective is an indefinitely extensible concept for much the same reason that *set* is. I suggested that lurking behind the argument from Russell's paradox was a much simpler intuition for why *set* is indefinitely extensible. The iterative conception of set is tied to a rule that I called generation by collection: the process of creating new

totalities out of the ones we have by collecting those totalities together.[16] The most important instance of this rule is the powerset operation, with which we form a new set from an old by collecting all the subsets of that set together. Something similar happens with perspectives. It is part of the essence of a perspective that it is a *partial* view on the world. This means that recognizing a perspective or collection of perspectives *as* perspectives requires us to take up some further perspective from which we can bring those perspectives into view as a kind of object, as incomplete views on the world. Without this, if we take up the perspective in question, we do not see it as a perspective, but simply as how the world is. That is, to see these things as mere perspectives on the world, rather than simply how the world is, we must reflect upon them from some more general vantage point—some other perspective. Thus it is part of our idea of a perspective that for something to be a perspective, there must be a another place from which we can assert this fact.

Call this process of standing back from a perspective so we might see it as such *generation by reflection*. It is the perspectival analogue of the powerset operation. Just as the powerset operation generates new sets by gathering the parts of an original set into a new set, generation by reflection creates a new perspective from which we look down upon the perspective we began with. History is rife with excellent examples. Macaulay tries to make sense of the of the world, or at least the space-time slice taking place in Britain from the assent of James II to the death of William of Orange. Years later Butterfield analyzes Macaulay's whiggish tendencies by studying the perspective that Macaulay brought to his history. Doing this requires Butterfield to take up a further perspective that allows him to situate Macaulay's history in its its political, social, and cultural context. Copernicus studied the solar system, Kuhn studied Copernicus's perspective on the solar system and the scientific context in which it was situated, and now more contemporary philosophers and historians of science study see Kuhn's work as offering a distinctive but incomplete perspective on the Copernican Revolution. In both of these examples understanding someone's

[16]See George Boolos (1989) for a discussion of the relationship between the iterative conception and the axioms of set theory. Talk of generation may strike some as either insidiously anti-realist or at least very metaphorical. Charles Parsons (1977) offers some advice on curing the iterative conception of these blemishes.

perspective *as* a perspective requires taking up a further perspective.

The process of taking up a new perspective for the sake of scrutinizing old ones is ubiquitous. Indeed, it is something we do to ourselves all the time. We hold up our attitudes, inclinations, and other opinions to scrutiny by understanding them as the product of a particular perspective on the world. Doing this means taking up a further perspective (but still *our* perspective) on these judgments.

Thus, to represent someone as having a perspective, a partial view of the world, requires us to take up a further perspective looking, as it were, over the original perspective's shoulder. This process leads to an ever-expanding circle of perspectives: to see A as a perspective, we must take up B, but to see B as a perspective, we must take up C, and so on.

Now someone could insist that this series goes on indefinitely, but that each perspective is nonetheless contained in some larger totality. This is claim is analogous to the thought that the hierarchy of sets goes on indefinitely but is still contained within a single maximal set.[17] Both of these claims fail for much the same reason. The assertion, "P is the totality of all perspectives" is self-defeating. To make such a judgment requires us to see the perspectives of P as just some partial views on the world among many others. But we cannot make this judgment from a perspective within P. Were we to occupy such a perspective we would not see this perspective, our own perspective, as an incomplete view of the world; we would see it as *how the world is*. If we want to say that P is a class of perspectives, we must take up a further perspective beyond P from which it is possible to make such a judgment. This means that we cannot say that P contains *all perspectives* (including the one I am occupying

[17]Bernard Williams appears to believe that all perspectives can be collected into a single totality. See his (1978/2005, pp. 65-6/49-50) and (1985, pp. 132-156). Williams says that beyond all perspectives on the world we must have an 'absolute' conception of the world, a conception that is not just another perspective on the world. We can then characterize the totality of perspectives in terms of this absolute conception, as the totality of things that are perspectives on this world. I think Williams's thought fails because his idea of an 'absolute conception' is an equivocation. The 'absolute conception' can be a genuine *conception*, a way of representing the world, in which case it is just another perspective, not a mysterious aperspectival representation. Or it can be just a way of gesturing at the final element in an infinite series formed by successively stripping away all the perspectival features of the world. But this isn't a "conception" of the world at all. At best it's a way of saying *that there is* a world standing beyond all our perspectives, but it is not a way of thinking about that world.

now) without giving the lie to that claim. This makes the statement self-defeating, and it shows that the generation of new perspectives by reflection cannot be stopped on pain of pragmatic contradiction (in the same way that the generation of new sets by collection cannot be stopped on pain of Russell's paradox). And from this we can conclude that *perspective* is indefinitely extensible.

3.7 How to be an open-ended constructivist

What does this mean for Socrates's Dilemma? Recall our dialectic. I tried to sketch a way in which constructivism could be made open-ended and thereby evade Socrates Subjectivity. This involved extending the construction procedure through time. The problem was that was that this construction was still Socrates Subjective because facts about our Goods were still constituted by a particular construction procedure encompassing the views of particular people. But this particular rebuttal gave us an idea. We can get our middle way if we deny the All-in-one Principle. What we need to do is insure that my critic's crucial assertion cannot be made, that there is no thing relative to which our construction procedure is Socrates Subjective—no thing for which something is Good because *it* says so. Making this thing infinitely large was a start, but not enough. We needed to make it unbounded, so that it cannot even be conceived of as a particular thing. Only that insures open-endedness while still denying Socrates Objectivity.

The claim that *perspective* is indefinitely extensible does that. The construction procedure for a Good can incorporate judgments from all perspectives. Such a construction would be absolutely open-ended because *perspective* is indefinitely extensible: we cannot form all perspectives into a single totality. And this means that our construction procedure is truly unbounded, not just large. Corollarily, the answer to Socrates's Subjectivity question is 'no': there is no class of attitudes, opinions, reactions, etc. or any construction procedure that takes such things as inputs for which we can say that something possesses a Good because those things say it does. More briefly, it is false that things are valuable, just, or justified in virtue of features of *us* because there could not possibly be any *us* large enough to make this proposition true.

At the same time, we still answer 'no' to Socrates's objectivity question. We deny that a thing possesses a Good because of the way it is in itself as all constructivists do. With this pair of answers, we have found a way around Socrates's Dilemma. We have a view of our Goods that is neither Socrates Objectivism nor Socrates Subjectivism. More importantly, *open-ended constructivism* offers an account on which our Goods have all the humanity that traditional constructivism offers, while also obtaining the special worldliness of transcendence.

3.8 Intensionality and reason

To some this proposal may look like a bit of logical prestidigitation. I may have shown how to evade the letter of both Socrates Objectivity and Socrates Subjectivity, but there seems to be something distressingly shysterly about my approach; I have evaded the problem without shedding any light on the underlying issues.

To counter this impression let me say a word about the intuitive underpinnings of open-ended constructivism. My claims about the indefinite extensibility of the concept *perspective* are just a way of making precise the more fundamental and more philosophical distinction introduced earlier. We can think of the transcendence of our Goods as a matter of our inquiries into those Goods being forever *pulled* toward some hidden outer reality, or we can think of them as being *pushed* from behind by an inborn urge to apply ever-increasing scrutiny. Philosophers have by and large assumed that the former is the only route to transcendence. But I recommend the latter.

A different way of making the same point relies on the distinction between extensional and intensional notions of functions, rules, and concepts. Suppose we are interested in the function $f(x) = x + 2$ over the natural numbers. We can characterize this function extensionally by identifying it with the ordered pairs $\{(1,3),(2,4),\ldots\}$. (Imagine that I have written down all of these pairs.) On this view, computing a function is like using a telephone directory: we look up the input in our set of ordered pairs and find our output. We can also understand the function intensionally. The function is a kind of black box that, in virtue of its inner workings, spits out $n+2$

whenever we hand it n. In simple cases these two conceptions coincide, and there is really no reason to mention the distinction. But sometimes the extensional characterization of a function gives out. Suppose we are interested in the identity function, $f(x) = x$. Further suppose we don't restrict the domain of this function to the natural numbers. In fact, we don't restrict it at all. We are interested in the identity function as it ranges over absolutely everything. But since sets are things and there is no set of all sets, there is no set of all things. Thus we cannot think about a set of ordered pairs (x, x) for absolutely all x, which, in turn, means that we cannot give an extensional characterization of our function. But we still have some understanding how this function works. If we are given 3, we return 3. If we are given a triangle we return the same triangle. If we are given Barbarossa, we return Barbarossa. Since this function cannot be characterized extensionally, this understanding must be intensional.

Open-ended constructivism offers an intensional conception of transcendence. The traditional, extensional approach to thinking about transcendence conceives of our inquiries in the same way that extensionalists conceive of the function $f(x) = x+2$. The scope of inquiry can be characterized by a set of perspectives on the world, and when we say that our Goods are transcendent, what we mean is that they are things in themselves lying beyond this set of all perspectives—beyond the 'extension' of inquiry. If we have this traditional picture in mind, then the transcendence of our Goods renders them Socrates Objective. But I am suggesting another way to understand transcendence. Our Goods are transcendent because *perspective* is indefinitely extensible. So when we say that inquiry ranges over all perspectives, we must understand 'all' intensionally. What does it mean to understand a quantifier intensionally? Look again at the case of sets. If we think that quantification necessarily involves collecting objects into a single domain, then we obviously cannot talk about all sets. On the other hand, we might suppose that quantification does not require collecting objects into a single domain. Instead we can quantify using a rule. When I say "include all sets in our hierarchy" I do not mean "include all $x \in S$ where S is the set of all sets", but rather "include s if s is a set", where this latter formulation functions as its own free-standing, intensional rule. If we adapt this way of talking about the question of transcendence, we can say that when I talk about 'all perspectives' I mean to enjoin

us to incorporate any perspective whatsoever into our construction procedure.[18]

This discussion allows us to introduce a trilemma in place of Socrates's Dilemma. A Good can be:

> *Extensionally transcendent* if its nature is not fixed by anything like a construction procedure because it has a nature independent of us and our attitudes, activities, etc.
>
> *Intensionally transcendent* if its nature is not fixed by anything like a construction procedure because any appropriate construction procedure is absolutely open-ended.
>
> *Non-transcendent* if its nature is fixed by something like a closed construction procedure.

Is this a *real* trilemma or, as some readers may suspect, a hair-splitting way of breaking one side of Socrates's Dilemma in two? ("Don't you *really* end up on one side of Socrates Dilemma or the other?" they ask.) If we think that this trilemma is a cosmetic splitting of Socrates's Dilemma, then we ought to be able to say on which side of the dilemma our middle option, intensional transcendence, naturally belongs. But there are two equally good assignments. If we think the essence of Socrates's Dilemma a distinction about constructivism versus realism, then we should assign intensional transcendence and non-transcendence to the same side. But if we think that what really matters is whether a Good is transcendent, then we should make the other partition and regard intensional transcendence as a variety of extensional transcendence whose outer nature is merely virtual. The fact that there are two equally natural ways to collapse my trilemma into Socrates's Dilemma shows that the trilemma marks out an interesting partition.

The partition also offers an attractive position. Extensional transcendence and non-transcendence present us with the problems we faced with Socrates's Dilemma.

[18]Compare the strategy employed by Vann McGee (2000) for quantifying over absolutely everything.

The first gives us worldliness without humanity, while the second gives us the opposite. But the middle option, intensional transcendence, gives us both: we get humanity in the same way that all constructivism does, and we get transcendence through the absolute open-endedness of our construction procedure.

For all I have said, one may still wonder *why* our Goods are intensionally transcendent. Saying that the concept *perspective* is indefinitely extensible may give us a mathematical demonstration of the point, but it does not explain very much. To get such an explanation we must unpack our guiding metaphor: the idea that transcendence follows from our inquiries being forever 'pushed' from behind.

I propose that the thing doing the pushing is reason. Or rather, a particular conception of reason: reason as a power for critical detachment, a capacity that we have to hold up our instincts, opinions, and activities to reflective scrutiny. On this conception reason is not a faculty for detecting new facts about the world, but something that applies scrutiny to our judgments. Elsewhere I argue that the way that reason scrutinizes is by asking whether these judgments would hold from other perspectives.[19] I may have an instinct to believe that the Sun orbits the Earth, but I can hold this proposed belief up to reflective scrutiny, and the way I do this is by comparing it to how I imagine things seem from perspectives beyond my own. Following an argument by Christine Korsgaard, I further suggest that this critical power of reason is the source of normative thought. For creatures without the critical power of reason, there is no question of what they *should* do, only of what they *will* do given their antecedent instincts and other attitudes. It is our special ability to detach ourselves from these attitudes that introduces the question of what we *should* do, above and beyond what we are inclined to.[20]

Another way to put this point is that reason so understood is what opens the questions that drive the Open Question argument. "I am inclined to do this thing, but *should* I do it?" and "I am naturally drawn to this thing, but is it *good*?" are questions put to us by reason when we detach ourselves from our inclinations and hold them up to scrutiny. And these questions are paradigmatically normative.

[19]See chapter 1.

[20]See Korsgaard (1996b, pp. 92ff) and Korsgaard (2009a).

What does this have to do with open-ended constructivism? Recall that the enduring openness of questions about our Goods was one way we characterized the transcendence of those Goods. However liberal we made our construction procedure, the question "this procedure calls this thing good, but *is* it good?" remains stubbornly open. If the critical power of reason is indeed the faculty that puts questions like these to us, then the nature of this power offers us an account of transcendence quite different from the one proffered by the Socrates Objectivist. Our Goods are transcendent because reason never ceases to scurtinizes our ideas about these Goods. And if I am right that reason applies this scrutiny by holding up our attitudes to more and more perspectives, then the indefinite extensibility of the concept *perspective* and the unceasingness of reason's scrutiny are two aspects of the same phenomenon. Moreover, if judgments made from various perspectives are the appropriate inputs of a construction procedure, then an open-ended construction procedure, like the one I have been recommending, will be a procedure driven by the demands of reason.

3.9 Kant on the unconditioned

Many readers are sure to notice the parallels between my account of the relationship between reason, normativity, and transcendence and Kant's account of the same. Kant held that pure reason is the source of normativity. It is pure reason's ability to bind us with laws that makes it possible to think about obligation.[21] Less well known is that the conception of reason I discuss here, reason as a source of endless criticism rather than new doctrines, was one explored by Kant in the first *Critique*.

> The greatest and perhaps only utility of all philosophy of pure reason is thus only negative. It does not serve for expansion, as an organon, but rather, as a discipline. It serves for the determination of boundaries, and instead of discovering truth, it has the silent merit of guarding against errors.[22]

[21] See the *Critique of Practical Reason*, 5:31.

[22] A796/B824. Translations are from Allen Wood and Paul Guyer's translation, Cambridge University Press, 1998.

Kant surely has this conception in mind when he tells us in the Transcendental Dialectic that "the proper principle of reason in general (in its logical use) is to find the unconditioned [...] which will complete its unity."[23] Save for the difference in nomenclature (my talk of perspectives versus Kant's talk of the 'conditioned'), Kant and I are making similar points. He says that reason is dissatisfied with judgments conditioned on perception, the understanding, and other parochialisms, and so it spurs us to strive for the unconditioned. Analogously, I see reason as dissatisfied with any circumscribed set of perspectives, any incomplete view of the world, and so spurring us to reach judgments that hold from absolutely all perspectives. The quarry of both hunts is the same: reason pushes us after an absolutely general, totally impartial, unvarnished view of the world.

But Kant doesn't think we should assume that the thing that reason seems to be driving us toward is real. Two paragraphs later he warns that "this need of reason [viz. the search for the unconditioned] has, through a misunderstanding, been mistaken for a transcendental principle of reason, a claim which over-hastily postulates an unlimited completeness in the series of conditions in the objects themselves."[24] Kant admonishes us not to infer from the unceasing demands of reason that there must be some transcendent thing (the unconditioned) that reason is striving after. In the same spirit, I admonish us not to think that transcendence must be understood extensionally, as a matter of some Socrates Objective thing standing beyond all perspectives. Both of these assumptions about transcendence are guilty of what Kant calls 'transcendental illusion'.

Kant's prescription is to see transcendence as "only a rule, prescribing a regress in the series of conditions for given appearances, in which regress is never allowed to stop with an absolutely unconditioned."[25] The dogmatist's mistake, then, is to reify the absolutely unconditioned and cast it as the subject matter of metaphysics. Analogously, I say that we understand the transcendence of our Goods intensionally, as a never-ending liberalization process driven by the scrutiny of reason. The Socrates

[23]A307-8/B364.
[24]A309/B366.
[25]A509/B537.

Objectivist makes the mistake of understanding inquiry as converging on some real, albeit infinitely distant, endpoint.

Finally, witness Kant's reliance on similar mathematical ideas to grope for a description of transcendence.

> I must always proceed empirically to a higher (more remote) member [of the series of conditions of appearances]. Thus the magnitude of the whole of appearances is never absolutely determined by that means. Therefore, one cannot say that this regress goes to infinity because this would [...] *determine* (though only negatively) the magnitude of the world prior to the regress. Accordingly we can say nothing at all about the magnitude of the world in itself, not even that there is the *regressus in infinitum* in it. Instead, we must only seek the concept of its magnitude according to the *rule* determining the empirical regress in it. But this rule says nothing more than that however far we may have come in the series of empirical conditions, we should never assume an absolute boundary, but rather we should subordinate every appearance as conditioned to another as its condition, and thus we must progress further to this condition; this is a *regressus in indefinitum* which, because it determines no magnitude in the object can be distinguished clearly enough from the regress *in infinitum*.[26]

In Kant's day there was an obvious rejoinder to this way of construing the rivalry between transcendental idealism and transcendental realism: a *regressus in indefinitum* can always be construed as a *regressus in infinitum*, just as Aristotle's potential infinities can always be construed as actual infinites. We need only think of the entire infinite sequence as a single totality.[27] The machinery I have introduced here allows us to regain the distinction Kant is groping after. Perspectives are not merely infinite in number. They cannot even be formed into a single totality, because the concept *perspective* is indefinitely extensible. For this reason the demands of reason can never be met.

[26]A519-20/B547-8.

[27]I think this kind of rejoinder lies behind the exasperated critique that Jonathan Bennett (1974, pp. 137ff) makes of these sections.

3.10 Open-ended constructivism in practice

The absolute open-endedness of a construction procedure makes it very hard to see how anything comes to possess a Good—how a society could be just, a loved one valuable, or an experimental method justified. If we avoid Socrates Subjectivity by insisting that our construction procedure never succeeds in fixing facts about Goods while also denying that these facts are fixed by things in themselves, then what sense can we make of trying to make our societies just and our methods justified, of our using normative concepts like 'just' and 'valuable' to regulate our personal and political behavior?

An open-ended constructivist can make sense of these practices in the following way. When we make a claim about a Good, we are relying on the outcome of a closed construction procedure. When we say that apartheid is unjust in the interest of effecting its end, we are basing that declaration on a construction procedure that does arrive at a definite answer, a procedure like Rawls's. But we should understand that closed procedure as a useful approximation of an open-ended construction procedure that fixes an *ideal* conception of justice. Rawls's construction is circumscribed in the range of perspectives it entertains, but we may nonetheless be justified in using it if we are convinced that it is the best choice among practicable approximations. Thus the way in which someone is justified in calling apartheid unjust is much different for the open-ended constructivist and the Socrates Objectivist. The latter thinks we are justified if, say, our judgment about apartheid is the result of a process reliably hooked up to the facts about justice. But I propose that our justification is more practical: we are justified in using a closed construction procedure to make a judgment if that procedure provides a suitable approximation for our ideal of justice given the context in which we are making this judgment.

We find the same arrangement in the use of models in science. The kinematic model of gases portrays gas molecules as massless point particles in random motion occasionally bumping into each other. This model embellishes many properties of gases: gas molecules are not point particles, they are not massless, their motion is not random, and their interactions are not, by and large, mechanical. But the kinematic

model is a good enough model to predict and explain the behavior of some gases in many circumstances. The model can succeed in these instances because the features it suppresses are not materially important to our target cases. And we prefer this model over one that does accommodate these features because it is vastly more cognitively tractable, and it produces a more elegant explanation.

Closed construction procedures like Rawls's have a status like that of the kinematic model of gases. Rawls's construction procedure leaves out some perspectives that could, in principle, matter to questions of justice. But for many of the questions we put to a theory of justice, we can have reasonable confidence that these omissions do not matter. And so we are justified in using Rawls's construction procedure, or another like it, to answer these kinds of questions, just as we use the kinematic model to answer certain questions about gases. An open-ended construction procedure for justice should therefore be regarded as something like a regulative ideal for our theories of justice, just as the idea of an empirically adequate and systematically organized theory is a regulative ideal for the natural sciences. First-order theorizing about ethics, politics, and epistemology amounts, then, to a kind of model-making. We build models of justice, value, and justification that strike a balance between two competing forces: the pull of the regulative ideal of a particular Good and the pragmatic force of the need for our model to be comprehensible to finite minds and implementable by human actors.

3.11 Conclusion

I began by describing two competing desiderata for a theory of normativity. We want our Goods to be at once humane and worldly. But these desiderata are hard to satisfy simultaneously. The incompatibility seems assured by Socrates's Dilemma: something is either Good in itself or Good because we make it so. The first horn of Socrates's dilemma promises to make our Goods worldly but not humane, while the second does the opposite. Many philosophers have tried to massage Socrates's Dilemma by constructing theories of normativity that are humane and worldly-ish or worldly and sort of humane. But these all come up short. The most ambitious goal

of this paper has been to join our two desiderata together by denying that Socrates's Dilemma is a true dilemma. Open-ended constructivism avoids Socrates Subjectivity by being absolutely open-ended. It avoids Socrates Objectivity by refusing to construe this open-endedness as a matter of our being pulled toward some transcendent normative reality. Instead, the view maintains that inquiries into our Goods are open-ended and the Goods themselves are transcendent because our own power of critical reason applies endless scrutiny.

Chapter 4

The theoretical and the practical

> You know my methods in such cases, Watson. I put myself in the man's place and, having first gauged his intelligence, I try to imagine how I should myself have proceeded under the same circumstances. In this case the matter was simplified by Brunton's intelligence being quite first-rate, so that it was unnecessary to make any allowance for the personal equation, as the astronomers have dubbed it. He knew that something valuable was concealed. He had spotted the place. He found that the stone which covered it was just too heavy for a man to move unaided. What would he do next?
> — Sherlock Holmes in 'The Adventure of the Musgrave Ritual'

4.1 The dogma of segregation

Philosophy is gripped by a dogma both pervasive and unspoken. It is the dogma that our theoretical projects, *limning the world*, and our practical projects, *thinking how to live*, are fundamentally independent of each other. I want to overthrow this dogma.

Of course, as the reader will be quick to point out, there are many ways that our theoretical and practical judgments appear to intermingle. Finding your house infested with radioactive cockroaches affects your plans to go inside, and your resolution to go to the laboratory affects what you believe about the intelligence of rhesus monkeys. But despite this, most philosophers still hold that this superficial frater-

nization masks a deeper separation between practical and theoretical judgments. We do indeed merge practical judgments about avoiding danger with theoretical judgments about what is dangerous to get the conclusion, "if you want to stay safe, you ought to avoid radioactive cockroaches." But this is kind of judgment is just the final packaging. Our dogma has it that these the judgments joined at this confluence must have been produced by something else, something further upstream. And these two somethings, our dogma says, are fundamentally separate: two independent faculties pursuing two independent goals by two independent means. Thus even if there we find coalescence between our theoretical and practical judgments, these judgments must have two separate origins: limning the world and thinking how to live.

This attitude of *segregationism* is widespread but largely implicit. And that makes it hard to articulate. Our first thought might be that the dogma is a version of the fact/value dichotomy. And it surely has *something* to do with the fact/value dichotomy, but I doubt this assimilation will make things any clearer. Everyone from Hume onward has offered a new gloss the dichotomy. The gap between fact and value is really a gap between two other things—'is' and 'ought', science and ethics, propositions and prescriptions, beliefs and desires. So saying that our dogma is a version of the fact/value dichotomy is not to say very much at all. The dogma is also orthogonal to our most familiar questions in metaethics: What is the semantic status of normative discourse? What is the metaphysics of value? What role do normative judgments play in our mental economy? These questions are related to our dogma, no doubt, but strictly speaking they all run askew of it. One can think that the judgment 'Chuzzlewit ought to give alms to the poor' is a *belief* representing a *fact* about *real* value in the world while still accepting the dogma. In that case, this judgment is formed theoretically. When we need to act, we merge it with some practical atttiude, like a desire to praise the charitable. One can also hold onto the dogma while taking quite the opposite attitude, while thinking that 'Chuzzlewit ought to give alms to the poor' expresses a desire or passion that corresponds to no facts whatsoever. In that case, we begin with a practical judgment about the duties of the rich to the poor, which then incorporates theoretical judgments about Chuzzlewit's wealth to produce a more specific passion about Chuzzlewit in particular. Thus our mad-dog realist and our

mad-dog anti-realist agree about the dogma; they just disagree about which side normative judgments fall under. The dogma is also orthogonal to the standard menu of views about moral psychology. For instance, one can be a Humean about motivation or an anti-Humean about the same and still believe in the dogma. The disagreement between these two foes is one about how to draw our dichotomy, not whether it exists.

We see then that this dogma is appreciably deeper than the familiar questions of ethics, metaethics, and moral psychology. For this reason it is hard to find it expressly formulated. We might have more success if we tried to find examples of the dogma animating particular positions.

Korsgaard. Perhaps the best example of the dogma is the dualism of standpoints that Christine Korsgaard finds in Kant:

> The deliberating agent, employing reason practically, views the world as it were from a noumenal standpoint, as an expression of the wills of God and other rational agents. [...] The theorizing spectator, on the other hand, views the world as phenomena, mechanistic and fully determined. The interests of morality demand a different conceptual organization of the world than those of theoretical explanation. Both interests are rational and legitimate. And it is important that neither standpoint is privileged over the other—each has its own territory. [...] These two standpoints give us two very different views of the world.[1]

So Korsgaard envisions two standpoints, "each [with] its own territory" and offering "very different views of the world". If we think that these standpoints produce their own, native judgments, then we have something very close to our dogma.

Gibbard. We find a similar schism in Allan Gibbard's distinction between planning how to live and describing the world. These activities, Gibbard tells us again and again, run in "parallel".[2] This is an apt metaphor because it captures both distin-

[1]Korsgaard (1989a, p. 173). The very same distinction appears in Korsgaard (1989b, pp. 377-8). Later in her career Korsgaard talks less about standpoints, but I think the underlying division survives in her denunciation of realism, for example in Korsgaard (2003).

[2]Gibbard (2003, pp. 66, 114-117, 181, 222-224, 235, 251-253).

guishing characteristics of Gibbard's program: the many similarities between planning and description, as well as their insuperable separation. Indeed, it is the whole point of Gibbard's program to show how similar planning and descriptive judgment are while punctiliously insisting on their segregation. This segregation is most stark when Gibbard talks about science. "Moore argued that oughts don't form a part of the natural world," Gibbard notes, "the picture I have sketched has the upshot that Moore was right. The scientific picture tells us why organisms like us would have questions whose answers can't be made a part of science." The scientific picture can't answer these questions because, "questions of what I ought to do and what it would be wrong to do or not to do aren't questions amenable to science. They are, I have been saying, questions of whether to help and of how to feel about not helping." If we take science as a stand-in for the theoretical, then Gibbard's rhetoric begins to sound like our dogma.[3]

Scanlon. The same dualism animates T. M. Scanlon's defense of his primitivism about reasons. To defuse naturalistic doubts about whether there could be such things as reasons, Scanlon proposes a picture of the relations between our projects rather like Carnap's. He suggests that we take a liberal attitude and brook many distinct and autonomous bodies of inquiry. These bodies of inquiry are unified by their subject matter and governed by stipulated rules. The domains he mentions are mathematics, science, morality, and practical reasoning. Scanlon uses this division of labor amongst different frameworks to rebuff charges that there is no room for reasons in our naturalistic conception of the world. Just as questions about whether there are electrons are not settled by mathematics, Scanlon says, questions about whether there is a reason for something are not answered by natural science. Rather, they are answered by our framework of practical reasoning. Thus Scanlon builds a methodological wall between many different kinds of endeavor, including one between the theoretical and practical.[4]

Shafer-Landau. Finally, we also our dogma in full-bore a moral realism like that of Russ Shafer-Landau:

[3]Gibbard (2008, pp. 17-8).

[4]See Lecture Two of Scanlon (2009).

> If moral realism is true, then there are facts about the moral status of situations—it is a fact, for instance, that torturing anyone for pleasure is wrong, and a fact that prudent and principled opposition to dictatorship is right. Do such facts necessarily supply us with reasons for action? It is important to see that realism *per se* is neutral on this question. Whether moral facts invariably supply reasons for action depends not on realism alone, but very importantly on which theory of practical reason one adopts. The truth of moral realism implies the existence of objectively correct answers to many moral questions. At issue here is whether we ought to take such answers as sound advice about how to behave. Here we have a set of facts—moral facts. And there we have a set of possible actions. Why should we let the first set influence decisions about the second?[5]

Thus Shafer-Landau thinks it an open question how our theoretical judgments, of which he thinks moral judgments are a species, interact at all with our practical judgments about what to do. In so doing he takes it for granted that these are entirely separate domains. What he, Korsgaard, Gibbard, and Scanlon have in common is this division, and whatever antagonism they come to is one about where items fit into this division.

These are just the more overt examples of the dogma. What seems to be central to each conception of the dogma is the independence of our practical and theoretical judgments, the thought that in some fundamental way the genesis, maintenance, and revision of each kind of judgment does not rely on the other. Our best strategy for refining the dogma would therefore be to sketch a criterion for this independence.

4.1.1 The independence of the theoretical and practical

The first candidate that leaps to mind is supervenience: the practical and the theoretical are independent if there is no interesting supervenience relationship between them. But this way of putting things places the cart before the horse. Pursuing this

[5]Shafer-Landau (2005, p. 165).

line of thought would require us to specify a supervenience base, and to do that we need to see what our best scientific theory tells us the world is like. But *that* will depend on the relationship between our theoretical judgments and our practical judgments. In other words, what we decide about our supervenience base will depend on whether or not our dogma is true. Suppose that our practical judgments are a part of our scientific theory—something that could happen if our dogma collapses—then it seems that supervenience holds trivially because our practical judgments are just supervening on themselves. On the other hand, suppose that practical judgments are completely independent of our scientific theory, as our dogma insists; then supervenience seems much less likely, but only for that reason. In either case supervenience is just a middleman between us and our real questions about whether there is meaningful interaction between theoretical and practical judgments.

In light of this, it would seem a better criterion would respect the relationship between the bodies of judgment we can use to specify a supervenience base—between what we might call *theories*. In these terms our dogma would have it that we have a 'theoretical' theory—our best scientific theory probably—that evolves in response to our theoretical demands and the available evidence. And we have an altogether separate 'practical' theory that evolves independently. We use our theoretical theory to answer questions about how the world is and our practical theory to answer questions about what to do. (I'm aware that 'theoretical theory' sounds like a pleonasm and 'practical theory' an oxymoron. But I think my meaning is clear.)

To make this idea precise we need to know how to individuate theories. The reigning story on this score turns on a notion already intimated by our talk of independence, the notion of a confirmation relationship. A judgment a is in a confirmation relationship with a judgment b if the acceptance or rejection of a could rationally affect our acceptance or rejection of b—that is: not merely *cause* the acceptance or rejection, but be a *reason* for it. Then we can say that a set of judgments S is closed under confirmation relationships just in case any judgment that is in a confirmation relationship with a member of S is also a member of S. In this spirit we might think of our theoretical and practical theories as the confirmation closure of some set of paradigmatic judgments.

But this won't quite work. Imagine there are some First Principles that affect the course of science but are not themselves confirmationally influenced by science—for instance the Principle of Sufficient Reason. We can even imagine them as a common bedrock influencing not only science but sundry other domains. It would be quite odd to say that these principles are a part of science; we certainly didn't discover them in the laboratory. It seems much better to say that they affect our scientific theories from the outside. The way to exclude such principles from our theory while still acknowledging their one-way influence is to require confirmation relationships to run in both directions. Thus, a judgment *a* is in a *mutual* confirmation relationship with a judgment *b* if the acceptance or rejection of *a* rationally affects our acceptance or rejection of *b* and *vice versa*, and a set *S* is closed under mutual confirmation just in case any judgment that is in a mutual confirmation relationship with a member of *S* is also a member of *S*.

These definitions gives us a practicable criterion of independence: two sets of judgments are independent if their respective closures under mutual confirmation relationships are disjoint.[6]

The ideas undergirding this criterion are the same ones standing behind Quine and Duhem's holism about confirmation. The first thought is that we should individuate our theories based on the work they do. If a collection of judgments are all working together toward some theoretical end, then those judgments are *thereby* parts of the same theory. The second thought is that judgments that are suspended in mutual confirmation relationships are doing their work *together*—they are explaining our target phenomenon as corporate bodies, not as individuals. Putting these two thoughts together, we get the idea that judgments are part of the same theory just in

[6]We need to be careful with this criterion. As stated, it could explode in such a way that *everything* is a part of a given theory because anything could in principle rationally affect our acceptance of another thing. In itself, explosion isn't a problem for the criterion. It jolly well *should* be an open possibility that all our endeavors live within a single theory. The problem arises only if explosion is somehow trivial. Whether this happens depends on our going notion of confirmation. There are surely some notions on which do not trivially lead to explosion—for instance whatever notion believers in rational intuition employ. So for anyone relying on this criterion, the proof of the pudding is in the eating. And so I beg the reader's patience in allowing me to show that the forms of confirmation *I* shall be concerned with are not the ones that yield trivial results about the union of two theories.

case they stand in a mutual confirmation relationship.

This my proposal for a criterion of independence produces a working statement of our dogma:

> **Segregation.** We have (at least) two theories, one we use to answer theoretical questions and one we use to answer practical questions. These theories can produce judgments that are combined into hybrid judgments, but the theories themselves are confirmationally independent.

I think that this is the dogma lurking behind the views propounded by Korsgaard, Gibbard, Scanlon, Shafer-Landau, and many others. It is the thesis I criticize here.

4.1.2 Why care about segregationism?

Before getting under way with my assault, let me say something about the significance of all this. The question of whether practical and theoretical judgments are part of the same or different holistic theories will strike some as painfully scholastic. But I think a great deal turns on it.

One issue is about prestige. Construing our thoughts about how to live as a wholly separate affair from theoretical inquiry can make them appear shabby. The writers I cited a moment ago are not guilty of this, but in less gentle hands the suggestion can make practical reason seem like a cosmic inconvenience. The practical accords we reach to avoid mayhem in the streets may be necessary, but they are not as lofty as knowledge about the depths of space or the center of the atom. We may be forced into making practical judgments, but only in the way we are forced to breathe and move our bowels—not because it is something worthwhile, but because we're put upon.

The second issue is about metaethics. I said at the outset that segregationism is strictly orthogonal to our traditional questions about the status of normative discourse. This is true, but the rejection of segregationism does make many metaethical views uncomfortable. Metaethics, like many branches of philosophy, is by and large about drawing distinctions and arguing over which side of these distinctions various

items rest. Many of these distinctions, then, depend on segregationism for their well-foundedness. So whether or not segregationism is true has seismic consequences.

Finally, the most important issue is about segregation itself. On the view propounded by segregationists, practical problems and theoretical problems are fundamentally different. Whether segregationists are right about all this is a considerable question in its own right.

My argument against the dogma of segregation bears on each of these. The failure of segregationism means that our practical judgments and our theoretical judgments are joined in a single holistic theory, a single web of judgment and principle. There will be, to echo Quine once again, some problems more practical and some more theoretical, but these will be differences of degree, not of kind. As a result, there will be no major difference of prestige. As a second result, some of the distinctions of metaethics disappear, or at least turn out shallower than we might have thought. Finally, and more positively, on my theory *thinking how to live* and *limning the true and ultimate structure of reality* are parts of the very same endeavor, and they cannot be put asunder.

Finally, to put my project into context, let me compare what I am up here to some other arguments with similar aims. In a recent book Hilary Putnam has argued that science and value are, in fact, tangled up with each other. Other writers point to thick concepts like 'cruel' and 'prudent' as examples of the same entanglement. And yet others insist that individual 'values', like the feminist value of being respectful of differences, and moral claims, like the wickedness of slavery, can be tested alongside more familiar hypotheses about the structure of the world.[7] I see my project as an order of magnitude more ambitious than these. Not only do I show that there *is* this kind of 'entanglement', I show that it is inevitable, I argue that its consequences are greater than we might imagine, and I explain *why* it is an essential feature of our lives.

[7]For examples of these approaches see Elgin (2007), Putnam (2002), Sturgeon (1988), and Anderson (2004).

4.2 How theories get married

My goal is to show that the practical and the theoretical are entangled in precisely the way that makes segregation impossible. My argument is an indispensability argument modeled on Quine's indispensability arguments. The most famous of these is about mathematics. Quine says that mathematical theories are just especially well-confirmed parts of a scientific theory, and to show this he adduces mutual confirmation relationships between our physical theory (traditionally conceived) and mathematics. First, he shows that the truths our physical theory expresses are influenced by the truths of differential geometry, topology, and, ultimately, set theory. Were the principles of these mathematical theories different, our picture of nature would be different too. And then *vice versa*: the question of which system of axioms for set theory we should adopt is settled in favor of *ZFC* plus Gödel's axiom of constructibility (rather than, say, *ZFC* plus the Continuum Hypothesis) because that system furnishes our physical theory with a simpler, more unified, less fanciful universe of sets. With confirmation relationships running in both directions, Quine concludes that mathematics and natural science inhabit the same holistic theory.[8]

I will pitch an argument rather like Quine's, only mine will be aimed not at segregation about mathematics and science, but at segregation between the theoretical and the practical. My case revolves around the special burdens imposed on our theories by the need to explain human action. Actions straddle the practical and theoretical worlds, and so however we see fit to fill in the details of our dogma, actions will lie on both sides of the distinction. It is this precarious situation that forces us to marry our theories. To explain an action we must explain two essential and inseparable facets of it: its physical features (how it influences the course of the physical world) and its intentional features (why it was performed, what it is meant to achieve, how reasoning brought it about). But, I shall argue, neither of our two theories on their own can furnish us with the materials to explain *both* aspects of action. Our practical theory

[8]The basic indispensability argument appears in Quine (1960, ch. 7). Quine's comments on constructibility are found in Quine (1990, pp. 94-5). I find the latter argument rather implausible because naturalizing higher set theory is a larger undertaking than Quine acknowledges. But no matter, I'm only interested in the form of the argument.

enables us to explain why a particular action had certain intentional features, and our theoretical theory can explain why it had certain physical features. But neither can do both. So if we want a full and comprehensive explanation—not just of the intentional aspects of an action and the physical aspects of the action, but of an action as a singular thing that is both intentional and physical—then we must use our two theories in concert. And this concerted use of our two theories toward a single explanatory end creates a mutual confirmation relationships, and this constitutes the union of our two theories.

Here is a brief example of the kind of argument I am offering, just to whet your appetite. Suppose Holmes sneaks a glance at Watson before quickly hiding his face behind a newspaper. To explain this action we must explain its mechanical aspects: the nerve impulses sent down Holmes's arm, the exchange of calcium ions in his muscle tissue, the elasticity of his ligaments. And we must also explain its intentional aspects: why it is the kind of action it is, why it makes sense for Holmes to do such a thing, whether it is done out of fear or flirtation, with an aim of concealment or arousal, displaying coquettishness or bashfulness. Without the former we've only explained a resolution, and without the latter we've only explained a movement. Thus our practical theory and our theoretical theory by themselves only offer half the explanation we need. The judgments proffered by our practical theory can explain the intentional features of the action (Holmes hid his face because he wanted to slyly attract Watson's eyes) but underdetermine the physical features. And the judgments our theoretical theory makes explain the physical features (Holmes hid his face because this nerve fired) but underdetermine the intentional. This pair of underdeterminations can't be tolerated. Because actions are single, unified things, not just some intentional stuff plopped on top of some physical stuff, we need a joint explanation, one that makes sense of the intentional and physical aspects of an action *together*. To explain Holmes's action, therefore, we need to use both theories, and we need to use them in tandem.

The next two sections make this argument in greater detail. They examine the two underdetermination claims I make. They ask, first, whether our practical theory can manage an explanation of action all by itself, and, second, whether our theoretical theory can do the same. After concluding that they cannot, I show that the result of

combining them into a single theory can. I conclude that the demands of explanation compel a merger between our two theories.

4.3 Just practical explanation?

Let us take as a working example the mysterious action at the heart of the story from which my epigraph comes. We have an agent—call him The Butler—who does lots of things. One thing he does is snoop through the family papers of his master, Lord Musgrave. He snoops through these documents because he wants to find a clue to the mystery of the Musgrave Ritual, which he thinks will lead him to a treasure hidden somewhere on the estate. One night The Butler sneaks out of bed in the hush of the night, descends to Lord Musgrave's study, and rummages through the family papers in hopes of finding a clue to the treasure. Lord Musgrave catches The Butler snooping and, perplexed by the whole affair, elects to consult his old college acquaintance, Sherlock Holmes. I begin by asking a more abstract version of the question Holmes might ask—

(4.1) Why does The Butler do what he does?

Our practical theory can offer us some semblance of an explanation. We can put ourselves in The Butler's shoes and use what we know about his beliefs and desires, hopes and dreams, demeanor and dispositions, to re-enact his deliberations. And this process of reconstruction can allow us to make sense of why he did what he did. He wanted the treasure because he lost all of his money at the track, because he wanted to marry a gentlewoman and needed a gentleman's income, because he was a collector in search of Charles I memorabilia. In other words we explainers can employ our own faculty of practical reason to decide which judgments would be apt given The Butler's state of mind, and we can then call on these judgments to explain what The Butler actually does.

But these sorts of considerations won't explain everything. They won't explain the physical features of The Butler's action—how The Butler propelled his body down the stairs, through the door, and into Musgrave's study. Most importantly,

they will not explain the transition from The Butler's deliberations to his physical movements in the world. Thus we see a gap between the conclusion of The Butler's deliberations (his setting out to do something, forming an intention, being primed for action, or however you might characterize it) and the physical changes in the world that he effects. Or, to revert to the segregationist's argot, a gap between the practical and theoretical worlds.

It should be obvious that this gap cannot be filled by more practical material. More facts about The Butler's beliefs, hopes, dreams, and ratiocinations won't move us any closer to an explanation of muscles contracting and blood pulsing—not, anyway, without some theoretical stuff to join up with.

And yet this is just what we need to explain The Butler's actions, to answer the question:

(4.2) Why does The Butler's deciding to solve the mystery of the Musgrave Ritual *bring about* The Butler's presence in Musgrave's study?

What this question asks is how The Butler's deliberations lead him to be in a particular physical state: hunched over a pile of papers, fingers tensed, eyes locked, pulse racing. We can put a finer point on the question by asking it contrastively:

(4.3) Why do The Butler's practical deliberations about Musgrave's family papers lead him to snooping rather than to some other state, say, firing a torpedo at the Bismarck or polishing the silver or becoming a buccaneer?

This is the sort of question, I propose, that our practical theory, our class of judgments about what to do, cannot answer by itself. And so it is the question on which our hopes for a purely practical explanation founder.

A popular answer to this challenge is a variation on a Davidsonian theme. The practical features of The Butler we have been dilating on—what we might call his motivating reasons—*cause* his action. Thus, we should say humdrum things like, "The Butler rummaged through Musgrave's papers because he wanted the treasure", and treat the word 'because' as a causal relation between reasons and actions.[9]

[9]Davidson (1963). If we don't like Davidson's particular version of this view, we might prefer a kind of indirect causation. For instance, The Butler's decision produces an intention, and intentions

But this kind of story does not bridge our gap between practical and theoretical worlds; it just gives us a new way to appreciate it. If we are going to explain The Butler's snooping by holding up a causal relationship between his motivating reasons and his action, then we need some means to discern the causal law connecting these relata. A one-off token-token causal connection may suffice to produce the relationship, but only something like a causal law, even if not a strict one, will explain it. But since the terminal relatum of this causal connection is a feature of the physical world, we will surely have to draw on our theoretical methods to pick out the explanatory law, to show that *this* law rather than *that* law explains The Butler's action. In other words, we need our theoretical theory, which makes sense of the physical world, to give us the causal explanation we are interested in.

A pessimistic thought we might lodge against this demand, also from Davidson, is that there simply are no such laws.[10] We see a similar idea in some recent readings of Kant, notably Tamar Schapiro's:

> [The agent] conceives of herself as a member of a "world of understanding," a world which is governed independently of natural law but which "contains the ground of the world of sense and so too of its laws." Because she must act within a sensible order, the free agent is in a position to exercise empirical causality. But because she identifies herself with her freedom, she sees herself as making things happen within a normatively structured world of which she is a co-author. The proper effect of free action, Kant writes, is "to furnish the sensible world, as a sensible nature in what concerns rational beings, with the form of a world of the understanding, that is, of a supersensible nature, though without infringing upon the mechanism of the former."[11]

cause actions.

[10]Davidson (1970).

[11]Schapiro (2001). The quotes from Kant are from, respectively, from the *Groundwork*, 4:453 and the *Critique of Practical Reason*, 5:43. The passage I quoted from Korsgaard earlier offers much the same picture of Kant.

On this story, action is something that takes place in the practical *cum* intelligible *cum* noumenal realm. This realm also 'grounds' the mechanics of the theoretical *cum* physical *cum* phenomenal world. But the process by which it does so, noumenal affection, is something we can never grasp. If this is how things work, then the demand I am placing on our explanations cannot be met.

But this is a very radical thought. Not only does it invoke the most metaphysically profligate understanding of transcendental idealism, and not only does it entail that we cannot explain actions, it denies that the relationship between us and the world around us is explicable. For Kant all such relationships are features of the phenomenal world, and so construing action as something in the noumenal realm makes our most ordinary activities deeply mysterious. I don't think we should be ready to say all this, at least not until we exhausted all other avenues.

And that brings us to the question of whether our theoretical theory can explain The Butler's action.

4.4 Just theoretical explanation?

We cannot explain the phenomenon of The Butler snooping with practical judgments alone because these judgments will not help us understand changes in the physical world. I doubt that many people outside a few dead-enders will disagree with this conclusion. I now come to the more controversial side of my claim, that we can't fully explain actions with our theoretical judgments either.

Let's suppose that what I have so far dubbed our 'theoretical theory' is more or less coextensional with the usual list of natural sciences: physics, chemistry, biology, human physiology. I claim that this body of method isn't up to providing us with an explanation of The Butler's action for reasons that parallel our practical theory's inadequacy. It produces all the judgments about calcium ion transfer, the elasticity of ligaments, and lock-picking that we need to explain the physical features of The Butler's snooping, but it cannot arrive at the facts we need to explain the intentional side of The Butler's action.

What do I mean by 'the intentional side of action'? For starters I mean all the

things that our practical theory was able to explain: why The Butler went to Musgrave's study instead of the butcher's shop, why snooping made sense given the kind of person he is, why he took the particular means he did, why his action was an instance of snooping and not tidying. More abstractly, we can say that the intentional features of an action are all those things that distinguish actions from mere movements.[12]

I say that our theoretical theory cannot adequately explain why our actions have these features. My argument comes in two steps.

> Step 1. Any explanation of the intentional aspects of an action must make reference to the deliberative episode that produced that action.
>
> Step 2. We can only explain this deliberative episode by citing judgments from our practical theory.

Therefore, we must employ our practical theory to explain the intentional aspects of action. I'll argue for each of these steps in turn.

4.4.1 For Step 1: The importance of deliberation

Suppose we say that The Butler's greed caused him to snoop through Musgrave's papers. There are two ways this might work. First, we could say that The Butler's greed directly moved him to snoop—without any intervening deliberation about what to do with this desire. This is not an action in our fullest sense of the word; it is a kind of affective mugging.[13] To see this event as an action it seems we have to say something about how The Butler's greed figured into his thinking about what to do. We might say that The Butler's greed enters into his deliberative soliloquy like so: "I need money, and snooping through Musgrave's papers is the best way available to get money. And I care about money more than everything else I can think of. So getting

[12]There is a glut of candidates for this distinction. See Millgram (2005) for a survey. Some say that an agent's actions are those movements that are somehow endorsed. Others say that actions have to be part of a practice of one sort or another. Yet others say that actions are those movements that provide self-knowledge. The list goes on.

[13]Davidson (1973) seems to agree. This case is analogous to his familiar example of a climber whose desire to drop his partner causes him to let go of the rope, only he lets go not because regarded this desire as a decisive reason to drop his partner, but because he was so shocked by his malevolence.

money is the thing to do." We might then say that The Butler's greed *caused* him to snoop because had The Butler not wanted money, he would not have snooped. But this counterfactual is true only modulo the course of The Butler's deliberations. It is true because The Butler regarded his want of money as a decisive reason to snoop.

All this means that the role played by The Butler's greed in his actions, insofar as they are full-fledged actions, is mediated by his deliberations. As such, any complete explanation of The Butler's actions must take note of these deliberations. They are what make The Butler's greed a reason for his actions, rather than the cause of movements.[14]

4.4.2 For step 2: Explaining deliberation

So we agree that any explanation of the intentional aspects of The Butler's action must say something about his deliberations. But what do we say? This is the crux of my argument so I will proceed a bit more slowly.

Our first question is whether we need to say anything at all. Can we just take the outcome of deliberation as given and explain The Butler's action in terms of that? We could do this, and we would get something of an explanation. It just wouldn't be a very good or very complete explanation. In a similar fashion we could explain the sea currents by pointing to the movements of the tides, which we regard as a brute fact. This is a fine explanation, but it could be bettered by also explaining the movements of the tides. The same, I suggest, goes for explaining action. Taking the outcome of The Butler's deliberations as a brute fact will get us some explanation, but there is certainly more to be had by asking why his deliberations turned out the way they did. So, let's ask: Why did The Butler's deliberations turn out as they did? How did he come to decide to snoop through Musgrave's papers?

I claim that there is but one way to answer these questions. We must figure out,

[14]The same point, but about a faculty of rationality instead of deliberation was made some years ago by Carl Hempel (1961) and more recently by Michael Smith (2009). It is true that in our day-to-day explanations of action, we typically cite motivating reasons as if they spoke for themselves, and in some instances it seems rather unlikely that an action was prefaced by a silent soliloquy. I can happily concede the existence of such actions, for as long as there are plenty of actions like the ones I am talking about, my argument goes through.

based on what we know about him, what it would make sense for The Butler to do. And judgments about what it would make sense to do, even conditional ones, are practical judgments. This means that we must rely on our practical theory to explain The Butler's deliberations. In other words, to make sense of The Butler's deliberations we must make practical judgments about what would be a sensible thing for a person like him to do in a situation like his.

My epigraph offers one way (though not the only way) to undertake this style of explanation. Holmes proposes that we simulate The Butler's reasoning:

> I put myself in the man's place and, having first gauged his intelligence, I try to imagine how I should myself have proceeded under the same circumstances. In this case the matter was simplified by [The Butler's] intelligence being quite first-rate, so that it was unnecessary to make any allowance for the personal equation, as the astronomers have dubbed it. He knew that something valuable was concealed. He had spotted the place. He found that the stone which covered it was just too heavy for a man to move unaided. What would he do next?

We take what we know about The Butler's beliefs and desires as inputs into our own practical methods, and we set about rehearsing his deliberations. If this process corresponds to what The Butler actually did, then we have our explanation: The Butler decided to ϕ because he is F, and deciding to ϕ is what it makes sense for someone to do when he is F (where F is a very large battery of attitudes, predilections, beliefs, and dispositions). If The Butler is indeed F and did decide to ϕ, then this is our explanation. Of course sometimes this doesn't work the first time. Perhaps we were wrong about The Butler's constitution. Perhaps his deliberations went differently than we imagined. Perhaps we made a mistake about which decision makes sense given what kind of a person he is. So we make these changes and start over again, and we keep trying until we get an account of The Butler's practical deliberation that both makes sense as an episode of practical reasoning and corresponds to what The Butler actually did. In any case, practical judgment is inevitably involved.

I think Plato first made this point in the *Phaedo* (97c-99a). Socrates observes that

we cannot explain his sitting in his jail cell just by pointing to his bones and sinews. We must mention his resolution to stay, and to understand *that* as more than a brute fact, we must appreciate the reasoning that lies behind it. We must understand what Socrates thinks about justice, mortality, and the good, and any appreciation of those views will necessarily engage our practical reason.

To reiterate the key question: why do we have to rely on *practical* judgments, ones about what to do, when explaining The Butler's deliberations?

Step back and consider the broad outlines of our explanatory project. We rely on three things in cobbling together this explanation: (i) the behavior The Butler exhibits, (ii) what we know antecedently about him, and (iii) the general demands of explanation—judgments about how (i) and (ii) must fit together for our account to make sense as an explanation. This final constraint is vital. There is a glut of stories about The Butler's deliberations that capture the data, but many of them are irredeemably bizarre. We need a standard of making sense to pick through all these empirically adequate accounts to find the one that is our best explanation.

This point is rooted in the nature of explanation. Explanation is the business of making sense of things, of making them intelligible. If we are trying to make sense of any part of the universe, be it The Butler's behavior or a distant galaxy, then we must assume that this thing can be made sense of—at least in a minimal way. So we must suppose some measure of intelligibility in our target phenomena, or our explanation will sink in the harbor. (An account that denies these virtues to our explanandum may still be true, but it will not be an explanation.) Innumerable criteria contribute to this evaluation: regularity, symmetry, simplicity, unity, and more romantic ones like elegance and beauty. Paying our respects to all this nuance would take us too far afield, so I propose we abbreviate it under the simple question of how much an explanation *makes sense*.[15] Thus to explain The Butler's deliberations we need to evaluate rival theories according to how much they make sense as accounts of practical deliberation, just as we need to evaluate rival theories of black holes according to how much they make sense as accounts of spacetime.

I think this fact about explanation is what Gibbard, Blackburn, and other critics

[15]For more on the role of evaluation in explanation see Peter Lipton (2004, ch. 9).

of practical explanation neglect. They suppose that an explanation of an action will be no more and no less than a chronicle of causes. And since *being the thing to do* is not the sort of property that fits nicely into the causal nexus, practical judgments won't show up in our explanations.[16]

But this is a naïve view of explanation. An explanation is not just a unorganized heap of causes, and if this rejoinder worked, it would be just as effective against the use of mathematics in explanation. Mathematical facts are not causally efficacious, and yet there they are in our explanations. The fact that a graph has an Eulerian circuit only if all vertices are all of even degree does not *cause* anything, but it does feature in our best explanation of the fact that a walk through Königsberg cannot involve crossing each bridge just once. Topology does not tie knots, but it does explain why this Hanover knot is so difficult to release.[17] In general, mathematics earns its keep in our theories not by denoting special mathematical entities play a causal role in producing our phenomena, but by offering resources that make our theories better explanations.

I suggest that practical judgments will enter into the process of explanation in much the same way. We have a set of potential explanations that are empirically adequate. Some of these are eliminated because there is no way that they will be consistent with what our theoretical theory tells us. But some are left over. We need a means by which to choose amongst the remainders. This selection will be based on our evaluation of which of these explanations makes the most sense of what we have seen.

The task of explaining an agent's theoretical reasoning provides a useful analogy. Suppose The Butler starts with a set of premises Γ and reasons his way to a conclusion δ. We could list the steps interjacent Γ and δ, but to understand these steps as *reasoning*, we need to show how they make sense as a deduction. And to do that we need to

[16]See Gibbard (2003, pp. 199ff) and Blackburn (1991). Many defenders of moral explanations make the mistake of fighting these critics on their own terms, of trying to find ways that moral terms can pick out causally efficacious parts of the world. For reasons I cannot fully enunciate here, I think this approach is misguided: to produce a proper defense of moral explanations we need to reject the claim that all explanation is causal.

[17]I take these examples from Christopher Pincock (2007) and Philip Kitcher (1989).

integrate them into some kind of deductive system. In other words, to determine which of several empirically adequate renditions of the reasoning from Γ to δ is the best one to ascribe to The Butler, we need to determine which one makes the most sense as a deduction. This does not mean that we must ascribe to The Butler the deductive system we think is the right one, but we should try as much as possible (sometimes it will be impossible, as in feminist bank teller cases) to see him as adhering to a system with particular logical virtues: consistency, soundness, and an aspiration for completeness. These same points hold, *mutatis mutandis*, for practical reasoning.

I am emphatically not saying that we must adopt the explanation that shows The Butler doing what makes the most sense *simpliciter*. So I am not saying that we cannot attribute mistakes to the agent. (What it makes sense to attribute and what it would make sense to do are, indeed, very different things.) I am saying that amongst the empirically adequate explanations we should choose the one that evaluates best.[18] This will very often be an explanation that is not terribly kind to the agent because the agent, in fact, did something stupid, irrational, or reckless. The point is that if we make the agent *too* stupid, irrational, and reckless we lose our ability to explain him. And the more we construe his deliberations as making sense—within the bounds of empirical adequacy—the better an explanation we will have of those deliberations. The analogy with physical explanation is once again telling. To say that we should prefer theories of the universe that are symmetrical and unified does not mean that we should automatically adopt the most symmetrical and unified theory we can possibly imagine—empty one-dimensional spacetime or something like that. But it does mean that if we let our theory get too ad hoc and disorganized, we risk its becoming something less than an explanation.

To sum up, we have to evaluate our candidate explanations in terms of how much sense they make of The Butler's deliberation, which in turn involves asking how much they make sense as accounts of deliberation. How do we do this?

It seems obvious enough: we have all these judgments about what it would make

[18]I should take that back just a bit. Sometimes we should be willing to sacrifice a smidgen of empirical adequacy if we can get a theory that is substantially more explanatory in return. How to make these sorts of trade-offs is, of course, a delicate question.

sense to do lying around. And these are not just categorical judgments, but conditional ones too. We have judgments about what it would make sense to do if you were keen on Charles I memorabilia and employed as a butler. We have judgments about what it would make sense to do if you believed that the key to the Musgrave Ritual was in the old family documents while yearning to solve that mystery. We have judgments about what it would make sense to do if you were in love with the scullery maid but too poor to marry her. Picking out which rival theory makes the most sense would seem, therefore, to just be a matter of applying these judgments.

But what kind of judgment are these? I think it is quite obvious that they are *practical judgments* and therefore members of our *practical theory*. For how could a judgment about what it would make sense to do be anything but a practical judgment? If we deny that this, then there is scarcely anything left that could be a practical judgment. And yet, I can imagine—and have encountered—a few strains of resistance to this claim.

The first tempting reply is to say that these judgments are just a way of stating our expectations. They are predictions, and so theoretical judgments. Saying it makes sense for The Butler to snoop is just like saying that it makes sense for a ball to fall when dropped, and this is just a way of saying that we should expect it to drop. But this suggestion is misguided because it relies on the wrong notion of making sense. It makes sense for a ball to fall when dropped because its falling is entailed by our best physical theory. And given that some account of The Butler's action is our best one, it would make sense for him to behave as it suggests. But *which account* of The Butler is best—most explanatory—is precisely our question. And to choose from amongst these alternatives we need a standard of making sense that is more than just a regurgitation of what is already in the candidate theories. We need a real evaluation.

The second line of objection accedes to my rebuff of this simple-minded construal of 'makes sense'. But it then goes on to accuse me of punning on the phrase, of conflating a theoretical evaluation of making sense, which is relevant to explanation, with an entirely different, practical evaluation endemic to deliberation. So, the objection goes, we do want an explanation of The Butler that makes sense, but this just means that we want it to possesses certain theoretical virtues, to be unified, sym-

metric, beautiful and all that. But that does not mean that the explanation must also possess some further practical virtues, that it must also make a special kind of practical sense. This point is eminently fair, but it is not a criticism of my proposal. I insist that we should introduce those evaluative standards necessary to providing the best explanation of an agent's actions—no more and no fewer. In practice I think *we will* end up introducing distinctively practical virtues (like whether an explanation sees an agent advancing her well-being, whether it is instrumentally sound, and so forth) because without them we wouldn't have much of an explanation. But that's just a conjecture. In principle there is no reason that the structural virtues we find physicists lusting after—symmetry, simplicity, beauty—won't be enough to single out a best explanation of The Butler's deliberations. My point is that in that instance, those virtues will be *per suum essentia* virtues of stories about what it makes sense to do. And thus they will be practical virtues. That's just what we are saying when we endorse them as criteria for our explanations. Indeed, I suggest that the thought that these might be the cardinal virtues of practical reason is the driving force behind Kant's universal law formulation of the Categorical Imperative. Secondly, respective of the charge of equivocation: these things have a nasty tendency to overgeneralize. If we try to nip my argument in the bud with a claim about the fragmentation of meaning for phrases like 'makes sense' and 'intelligible', then there's nothing to stop us from saying the same thing about the standards of evaluation amongst the different sciences. We might suggest that broadly the same intelligibility criteria are in order for physical and biological theories because we want both kind of theory to make sense. But against this thought another person might protest that we are punning: there are distinctive kinds of making sense for biology and physics, so we should not expect a unified set of explanatory virtues for biology and physics. This is a ridiculous thing to say, of course, but there seems to be just as much reason to splinter the meaning of 'makes sense' at this point as any other. So the the uniformity of standards of intelligibility across our various explanatory targets—including action—strikes me as an eminently reasonable working assumption, even if not an unassailable postulate.

The third line of resistance is a more radical version of the second. It bites the bullet and says that judgments about *what it makes sense to do* are theoretical. The only

judgments that are properly practical are those about *what to do*. The words 'makes sense' make all the difference. On this view, our explainer has a body of judgment about what it would make sense to do, which she uses to explain The Butler's behavior. She also has a set of practical judgments about what she herself should do. And these two bodies of judgment are completely independent. This arrangement seems logically consistent, but in practice it is clearly unstable. It all comes crashing down the moment our explainer realizes that she is a member of the world, a person whose actions must explained alongside everyone else's. If this person keeps these two sets of judgments strictly separate, then there is a possibility she could say, in the same breath, 'it makes no sense for me to snoop' and 'snooping is the thing for me to do'. This is a blatant contradiction. There may be ways to wriggle out of the noose, but I doubt any of them are pretty. This epiphany about one's place in the world is a transformative moment. The awareness that we are an agent like any other shows that we cannot hope to keep separate our *truly* practical judgments about what *we* shall do and our pseudo-practical theoretical judgments about what it would make sense for *others* to do. We are, after all, just like the others. Therefore, unless we are invidiously self-involved or fatuously self-unaware, there is just one set of judgments about what to do and what it would make sense to do that perform double duty: we use them to guide our own actions and to evaluate potential explanations of others.[19]

None of these objections to my suggestion are successful. We therefore have no choice but to accept that these crucial evaluations are what they seem to be: practical judgments. So I conclude that we must use our practical theory if we are to give a complete explanation of The Butler's deliberations. And this means that our theoretical theory cannot, by itself, explain the intentional features of his actions. Before putting this point together with our previous conclusion about the inadequacy of our practical theory, let me make two important clarifications.

[19]The most insightful accounts of I what I called the 'transformative moment' come from phenomenologists, notably Sartre in *Being and Nothingness*, Book Three, §1.IV.

4.4.3 Simulation and separation

My mention of Sherlock Holmes and his practice of putting himself in another man's shoes may give the impression that I think we can only make the practical judgments necessary to explaining The Butler's deliberations by *simulating* those deliberations, that we have to imaginatively project ourselves into his place and reason about what to do as if we were he. This is an idea with a pedigree. Writers as far back as Wilhelm Dilthey and R. G. Collingwood have held that understanding another's actions requires a kind of simulation. And they have thought, contrary to my own conclusions, that this point distinguishes empathetic *Geisteswissenschaften* from nomic *Naturwissenschaften*.

As interesting as this suggestion is, it is not mine. Simulation is one way to get at the requisite practical judgments, but it is not the only way. We can proceed intensionally, by putting ourself in The Butler's shoes and using our own practical judgments to produce the theory that makes the most sense, which we then compare to the hard data. Or we could proceed extensionally, by taking all empirically adequate theories and ranking them in terms of how much sense they make. This dualism of approach once again mirrors what we see in the physical sciences. We can start with a pristine mathematical model and adjust it as recalcitrant data come in, or we can start with a mess of data and try to build a model around it. In practice our explanations, of physical systems and people, no doubt involve a little bit of both.

So simulation is not the royal road to explanations of actions. But even if it were, I still see no good reason for the kind of separation that Dilthey and Collingwood propose. We may need simulation to discover the principles under which we organize our explanation of The Butler's deliberations. But we need simulation for lots of other things too—there is even a literature suggesting that our understanding of causation requires a kind of personal simulation.[20] And we might not be able to formulate universal laws of behavior, but the same is arguably true of all the special sciences. As a general matter, the temptation to think that there is a schism in the explanatory aims of our projects grows out of a romanticized picture of how explanation and inquiry

[20]I am thinking of what is called 'manipulationism' or 'interventionism' about causation, the view on which our notion of causation is inextricably tied to our conception of action.

work, one that fails to appreciate just how messy even our canonical examples are.

4.4.4 Axioms of practical reason

Finally, I want to distinguish my milquetoast claim that we must employ our own practical theory, whatever it may be, in explaining an agent's action from proposals that we must use some *particular* axiom of practical reason as a rubric in our explanations.

The most popular suggestion is "*nihil appetimus, nisi sub ratione boni*", that nothing is desired except under the guise of the good. We find this idea in first sentence of the *Nicomachean Ethics* and in the crucial step in Kant's argument for the Formula of Humanity. More recently, David Lewis and Davidson have included 'rationality constraints' in their systems of radical interpretation that require us to make our target subject as rational as possible, in Davidson's words to make him "believer of the true and a lover of the good". And Jonathan Dancy has insisted that when we explain an action, the considerations we point to as motivating that action must themselves be capable of being *good reasons* for the action. Of course some writers reject this kind of constraint. Chief among them are David Velleman and Kieran Setiya, who argue that people act not under the guise of the good, but, roughly, to understand themselves.[21]

I am officially neutral in this debate. I do not endorse any particular views about the shape or content of our practical theory or how it it enters into our explanations. My claim is a generic one. We must rely on our practical theory, whatever form it may take, in our explanations of an agent's deliberations. My argument also works rather differently from these briefs for particular axioms. They tend to start with observations about what practical reason is like from the inside and insist that any explanation of an agent's deliberations must reflect those observations. What I am urging, though, is quite different. It is that explanation itself requires us to make judgments about what makes sense, and when we are explaining actions these judgments will be practical ones *per se*. Thus the use of practical reason arises organically from our explanatory needs, not from any doctrines about the logic of deliberation.

[21]See Davidson (1970), Lewis (1974), Dancy (2000, ch. 5), Velleman (1992), and Setiya (2007, ch. 1).

4.4.5 Summing up

In the way of a recapitulation, let me field one final objection. It amounts to an incredulous stare:

> How could a practical judgment about what to do be at all relevant to an explanation? Isn't it just the *wrong kind of thing* (it has the wrong 'direction of fit') to be explanatory?

The answer goes back to the criticism I lodged against Blackburn and Gibbard's naïve conception of explanation. If explanation were just a matter of joining events together into a lattice of causal relations, then we could use this riposte to undercut the explanatory potential of any number of things. But it's not. Explanation is a charge to make sense of some target phenomena, and to judge something as making more or less sense is an evaluation. Sometimes our target phenomenon will be very alien, and this evaluation will turn on abstract criteria like unity and symmetry. But sometimes the phenomenon will be something more intimate, something that we ourselves partake in. In these cases we should expect our evaluation to be more nuanced and more specific. This is what we see when our target phenomenon is action. Because we ourselves are constantly going back and forth between reasoning about what to do and coming to grips with the doings of others, we should expect these two projects to affect each other. We should expect the evaluations concomitant to our explanations of others' deliberations to be entangled with the evaluations we make in our own deliberations because they are evaluations of the same thing. To suppose otherwise is to see oneself as radically separate from the rest of the world. Thus practical judgments affect our explanations not because they describe some bit of reality (they couldn't) but because they lead us to those explanations that make sense.[22]

[22]Caveat emptor: I have not argued that practical judgments will make their way into the *content* of our explanations (as opposed to the materials we use to produce and express that content). All I have said is that our methods for producing explanations must employ both practical and theoretical judgments. In recent years this issue has come alive, as some writers have tried to insist that explanations using mathematics (*inter alia*) do not actually have mathematical content. Instead mathematics is just a useful tool for specifying the content of the explanation. An analogous claim could be targeted at my thesis here. For my part, I think this program is wrong-headed: we should construe the content

4.5 Our theories married

The last two sections showed that neither our theoretical theory nor our practical theory include all the judgments we need to explain The Butler's action. Our practical theory cannot explain how the physical features of an action came to be. And our theoretical theory cannot explain why an agent decided to do as he did.

What can we say about this double failure? Some readers may be tempted to keep our two methods segregated and explain the physical and intentional aspects of our action in parallel. Explain all the intentional features with our practical theory and all the physical features with our theoretical theory. Two independent phenomena, two independent explanations, two parallel aspects of action. This, I think, is the view of action endorsed by Korsgaard and Schapiro's Kant. Other examples are not hard to find.

Nonetheless, this Panglossian delusion of pre-established harmony cannot be abided. In point of fact, an action is not two separate phenomena. It is something that is both intentional and physical, something that originates in deliberations about what to do and ends with a change in the world. The intentional *sans* physical is just an attempt, and the physical *sans* intentional is just a movement. What we must ultimately explain is the unity of action: how these intentional and physical features coalesce into a single thing.

This observation is the crux of my argument against segregationism. We cannot keep our theoretical and practical theories separate because there is a particular phenomenon where they must come together, and that phenomenon is human action. As long as our methods work along parallel tracks, this nexus of the practical and the theoretical will be mysterious. So if we are to explain action as a single, unified thing, we need to merge our two methods. We need to offer a single explanation that sheds light on the essential union of physical and intentional in each of our actions.

I realize this grand program sounds mysterious, so here are some examples of how it might work. What we see in the following examples is a process of iterative

of our explanations as broadly as possible. That said, right now I am only interested in explanation as it affects confirmation, not in the contents of explanation themselves, so I am officially indifferent to this debate.

refinement. We start with provisional hypotheses about the intentional and physical aspects of an agent's action and then oscillate between our two methods, gradually refining these hypotheses until we reach a point of equilibrium on which they fit together to form a coherent explanation of the whole.

Case 1: The Kick. Suppose we observe The Butler kick Lord Musgrave in the shin. Our first thought will be that The Butler kicked Musgrave to hurt him. But we know antecedently that The Butler did not bear any ill will toward Musgrave, so The Butler's kicking Musgrave to hurt him wouldn't make sense—a conclusion we reach by deploying our own practical theory. The rejection of our initial hypothesis on the grounds that it fails as an intentional explanation then leads us to an alternative hypothesis about the physical features of The Butler's kick. We might think that it was caused by his patellar reflex, and so was quite involuntary. But we check The Butler's knee with our best scientific methods and find that it was not a reflex. So we offer another hypothesis: perhaps The Butler kicked Musgrave out of play. We take out our practical theory once more and hypothesize that this would make sense if he wanted to flirt with Musgrave; life in the English countryside can be frightfully boring after all. But we return to our theoretical theory and find that The Butler's kick was far too ferocious for it be flirtation. (Another practical conclusion.) So we start over again and hypothesize that The Butler meant to disable Musgrave in order to escape somewhere. But this raises further practical questions. And so on.

Case 2: The Prowl. Now suppose we observe The Butler sneaking into Musgrave's study. We are poised to explain this by saying that he wanted to rob Musgrave to pay off his gambling debts. This does make some sense given what we know about The Butler. But this is hard to square with some of the physical aspects of The Butler's action, like his going for old family documents instead of money. So we go back and revise our hypothesis: The Butler snoops through Musgrave's papers out of a prurient fascination with the lineage of well-bred men. But when we check this hypothesis against the facts gleaned from our theoretical methods we are disappointed, for we find that The Butler's skin conductance doesn't increase while he paws through these documents—usually a sure sign of arousal. So we return once more to our practical theory and try to think of a scenario under which it would make sense for The Butler

to snoop through Musgrave's documents in absence of any intrinsic interest in them. We then come up with the thought that he hopes to find clues to the location of the treasure. And so on.

These attempts at explanation are all very pedestrian, and they say precious little about how exactly the practical mind might be connected to the physical world. But that wasn't my intention. The examples show how we oscillate between our two methods to reach a satisfactory explanation. We offer an intentional account, revise it in the face of some recalcitrant judgment from our theoretical method, find this revised judgment wanting by the lights of our intentional method, and so revise it again. We repeat this process, going back and forth, until our intentional and physical explanations are brought into equilibrium with each other. This is what I mean when I say that we must use our methods in concert, and this is how we must ultimately explain an action.

Some readers will no doubt be disappointed by this strategy. I have made a terrific fuss about the gap between deliberation and the physical world, but it is hard to see how this mealy-mouthed equilibrium story will fill that gap. At least theories like Davidson and Kant's offer us some story along those lines, even if not a proper explanation. My proposal appears too anodyne for even that.

But this is the wrong way to think about what I am offering. I am not suggesting a way to fill the great gap. I am showing how we might think about the relationship between the theoretical and practicals worlds in such a way that the gap never arises. We should not think of our deliberations, which are governed by our practical theory, and the external world, which is described by our theoretical theory, as two closed systems that must be bridged by some special relations. Instead, we should think of those two systems as already integrated with each other in our everyday inquiries, integrated in such a way that this gap never arises. This is what my story about explanation is supposed to accomplish. We have very ordinary explanations of behavior that oscillate between a whole bevy of practical and theoretical judgments. And this, in effect, rubs away any gap between the two systems. This is all less exciting than the thunderous reduction that many philosophers instinctively reach for, but sometimes the best answer is the less audacious one.

4.6 Integration

Recall our confirmation closure criterion for theory inclusion. Because scientific theories are confirmed and denied as internally connected wholes, not piece by piece, we ought to construe their borders so as to include anything and everything that is involved with their overall explanatory project.

Thus if two classes of judgment are in a mutual confirmation relationship, then they are part of the same overall theory. This is what Quine proposed about the physical sciences and mathematics: he showed that our mathematical judgments affect which physical judgments we accept and *vice versa*. We now see the same relationship for our hitherto separate theoretical and practical theories.

Our aim is to produce the best overall explanation of actions like The Butler's. (In saying this I am assuming the validity of a form of inference to the best explanation). But as we saw, to get this best explanation we need our theoretical theory and practical theory to 'work together'. In practice this means that each kind of theory must be revisable in the face of what the other says: we put forward a hypothesis about The Butler's action that involves a theoretical judgment and a practical judgment, we find it wanting, so we go back and revise our theoretical theory to produce a new judgment. But this new judgment doesn't fit with our old practical judgment, so we revise our practical theory to produce a new practical judgment. And so on. What is important is that each theory influences our revisions of the other, albeit mediately through their mutual goal of producing the best overall explanation.

Quine often talks about how we can accommodate disturbances in the scientific picture by making 'compensatory adjustments' elsewhere in that picture. In these terms, my argument shows that making sense of action requires our practical and theoretical judgments to be joined in a single network of compensatory adjustment.

This process, if we are lucky, produces a single, coherent explanation of an action. If we did not have this mutual confirmation relationship, we would instead have two disconnected explanations—one of mere deliberation and another of mere movement. The fact that we must countenance these mutual confirmation relationships to get our best explanation of an action is, I propose, all the reason we need to

see our two theories as parts of a larger whole.

This, then, is my thesis about practical judgments:

> **Integration.** Practical judgments and theoretical judgments are part of the same holistic theory.

If we take this whole grand method to be science, and practical judgments to be judgments about what to do, then we can dress our thesis in slightly more ostentatious regalia. We can proclaim that our scientific methods can, in principle, settle what we ought to do.

4.7 Conclusion

With my attack against segregationism completed, we are in a better position to evaluate its import. First of all, the conclusion has obvious implications for how we conceive of our lives and the world around us. There is a temptation to think that finding out about the world and leading our lives are two rather different things, that it is just a coincidence that human beings happen to do both. This natural view, I have tried to argue, is an illusion. There is one project encompassing both our theoretical inquiries and our deliberations about how to live.

Then there are the questions of metaethics. It is true, as I remarked at the outset, that the division between practical and theoretical is strictly orthogonal to the paradigmatic questions of metaethics. But the collapse of this dichotomy does affect our answers to these questions because it affects what kind of distinctions we can draw. Chief among these effects is the impossibility of a metaphysical asymmetry between theoretical and practical judgments. On the big questions about realism, practical judgments stand and fall with our theoretical judgments. If we want to be rabid realists about science, then we ought to be rabid realists about the practical too. And if this stance proves uncomfortable, then that is a problem for scientific realism, not for practical judgments. Likewise, if we are dead set on being constructivists about practical discourse, then we have to take the same view for all of science.

More pointedly, insofar as the metaethical views I canvassed at the beginning of the paper are really committed to this dichotomy, they are unsustainable. So, insofar

as Korsgaard's view is built on the assumption that there are two distinct standpoints on the world, it is untenable. Of course, we commonly talk about some standpoints being more practical (planning an invasion of France) and others more theoretical (making a map of France), but these differences of degree are not enough for Korsgaard's program. Similarly, Scanlon's variations on a theme by Carnap are untenable for the very same reason that the original was: our practices do not respect the divisions that Scanlon posits. We do talk about some questions belonging to physics and others to mathematics, but this kind of talk is not enough to support Carnap and Scanlon's lofty ambitions. By the same token, expressivism is out as an *a priori* claim, but if it is understood as an empirical conjecture—comparable to the idea that mathematics and language are handled by different mental modules—then it is quite compatible with all I have said. But in that case I don't see how it can do all the work that some of its advocates have promised.

In general, deep divisions between theory and practice are unsustainable, but shallow ones very well may persevere. After all, there is obviously a difference between the Ten Commandments and the *CRC Handbook of Chemistry and Physics*, between planning a get-away and looking through a microscope, between belief and desire, between love and curiosity. My point is just that none of these distinctions can be the titanic one that many have supposed. It will be more like the distinction between the different branches of science than different domains of reality. There will be a difference, but it will be rough, hazy, sometimes arbitrary, other times invidious, and, above all, quite incapable of bearing any philosophical weight.[23]

[23]What I am offering, then, is in the same spirit as Hilary Putnam's (1962) subtle rejection of the analytic/synthetic distinction: there is some distinction, but it isn't philosophically interesting or useful.

References

Anderson, E.: 2004, Uses of value judgments in science, *Hypatia* **19**(1), 1–24.

Bennett, J.: 1974, *Kant's Dialectic*, New York: Cambridge University Press.

Blackburn, S.: 1991, Just causes. Reprinted in Blackburn (1993).

Blackburn, S.: 1993, *Essays on Quasi-realism*, New York: Oxford University Press.

BonJour, L.: 1998, *In Defense of Pure Reason*, New York: Cambridge University Press.

Boolos, G.: 1989, Iteration again. Reprinted in Boolos (1999).

Boolos, G.: 1999, *Logic, Logic, and Logic*, Cambridge: Harvard University Press.

Boyd, R.: 2003, Finite beings, finite goods: The semantics, metaphysics and ethics of naturalist consequentialism, *Philosophy and Phenomenological Research* **66/67**(3/4), 505–553/24–47. In two parts.

Brandom, R.: 1998, *Making It Explicit*, Cambridge: Harvard University Press.

Child, W.: 1996, *Causality, Interpretation, and the Mind*, New York: Oxford University Press.

Dancy, J.: 2000, *Practical Reality*, New York: Oxford University Press.

Darden, L. and Maull, N.: 1977, Interfield theories, *Philosophy of Science* **44**(1), 43–64.

Daston, L.: 1992, Objectivity and the escape from perspective, *Social Studies of Science* **22**(4), 597–618.

Daston, L. and Galison, P.: 2007, *Objectivity*, New York: Zone Books.

Davidson, D.: 1963, Actions, events, and causes. Reprinted in Davidson (1980).

Davidson, D.: 1970, Mental events. Reprinted in Davidson (1980).

Davidson, D.: 1973, Freedom to act. Reprinted in Davidson (1980).

Davidson, D.: 1980, *Essays on Actions and Events*, New York: Oxford University Press.

Davidson, D.: 1984, *Inquiries into Truth and Interpretation*, New York: Oxford University Press.

Dennett, D.: 1987, *The Intentional Stance*, Cambridge: MIT Press.

Dennett, D.: 1991, Real patterns, *Journal of Philosophy* **88**(1), 27–51.

Dummett, M.: 1993, *Seas of Language*, New York: Oxford University Press.

Elgin, C. Z.: 1996, *Considered Judgment*, Princeton: Princeton University Press.

Elgin, C. Z.: 2007, The fusion of fact and value, *Iride* **20**, 83–101.

Enoch, D.: 2006, Agency, shmagency: why normativity won't come from what is constitutive of agency, *Philosophical Review* **115**(2), 169–198.

Ferrero, L.: 2009, Constitutivism and the inescapability of agency, *in* R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 4*, pp. 303–333.

Flanagan, O., Sarkissian, H. and Wong, D.: 2007, Naturalizing ethics, *in* W. Sinnott-Armstrong (ed.), *Moral Psychology, Vol. 1*, Cambridge: MIT Press.

Foot, P.: 2003, *Natural Goodness*, New York: Oxford University Press.

Frankfurt, H.: 1969, Alternate possibilities and moral responsibility. Reprinted in Frankfurt (1988).

Frankfurt, H.: 1988, *The Importance of What We Care About*, New York: Cambridge University Press.

Friedman, M.: 1974, Explanation and scientific understanding, *Journal of Philosophy* **71**(1), 5–19.

Friedman, M.: 1991, Regulative and constitutive, *Southern Journal of Philosophy* **30**.

Friedman, M.: 2002, Kant, Kuhn, and the rationality of science, *Philosophy of Science* **69**(2), 171–190.

Gibbard, A.: 2003, *Thinking How to Live*, Cambridge: Harvard University Press.

Gibbard, A.: 2008, *Reconciling Our Aims*, New York: Oxford University Press.

Goodman, N.: 1955, *Fact, Fiction, and Forecast*, Cambridge: Harvard University Press.

Goodman, N.: 1976, *Languages of Art*, Indianapolis: Hackett Publishing.

Gould, S. J.: 2002, *Rocks of Ages: Science and Religion in the Fullness of Life*, New York: Ballantine Books.

Habermas, J.: 1998, *Inclusion of the Other: Studies in Political Theory*, Cambridge, MA: MIT Press.

Hempel, C. G.: 1961, Rational action, *Proceedings and Addresses of the American Philosophical Association* **35**, 5–23.

Hempel, C. G. and Oppenheim, P.: 1948, Studies in the logic of explanation, *Philosophy of Science* **15**(2), 135–175.

Horkheimer, M.: 1933, Materialism and morality. Reprinted in Horkheimer (1993).

Horkheimer, M.: 1937, Traditional and critical theory. Reprinted in Horkheimer (1972).

Horkheimer, M.: 1972, *Critical Theory: Selected Essays*, New York: Seabury Press.

Horkheimer, M.: 1993, *Between Philosophy and Social Science*, Cambridge: MIT Press.

Kitcher, P.: 1989, Explanatory unification and the causal structure of the world, *in* P. Kitcher and W. Salmon (eds), *Minnesota Studies in the Philosophy of Science Vol. 8: Scientific Explanation*, Minneapolis: University of Minnesota Press.

Kornblith, H.: 2002, *Knowledge and Its Place in Nature*, New York: Oxford University Press.

Korsgaard, C. M.: 1989a, Morality as freedom. Reprinted in Korsgaard (1996a).

Korsgaard, C. M.: 1989b, Personal identity and the unity of agency: a reply to Parfit. Reprinted in Korsgaard (1996a).

Korsgaard, C. M.: 1996a, *Creating the Kingdom of Ends*, New York: Cambridge University Press.

Korsgaard, C. M.: 1996b, *The Sources of Normativity*, New York: Cambridge University Press.

Korsgaard, C. M.: 2003, Realism and constructivism in twentieth-century moral philosophy. Reprinted in Korsgaard (2008).

Korsgaard, C. M.: 2008, *The Constitution of Agency*, New York: Oxford University Press.

Korsgaard, C. M.: 2009a, The activity of reason, *Proceedings and Addresses of the American Philosophical Association* **83**(2).

Korsgaard, C. M.: 2009b, *Self-Constitution: Agency, Identity, and Integrity*, New York: Oxford University Press.

Koyré, A.: 1957, *From the Closed World to the Infinite Universe*, Baltimore: The Johns Hopkins University Press, 1968.

Lange, M.: 2009, *Laws and Lawmakers: Science, Metaphysics, and the Laws of Nature*, New York: Oxford University Press.

Lewis, D. K.: 1973, *Counterfactuals*, Cambridge: Basil Blackwell.

Lewis, D. K.: 1974, Radical interpretation. Reprinted in Lewis (1983).

Lewis, D. K.: 1983, *Philosophical Papers Volume I*, New York: Oxford University Press.

Lipton, P.: 2004, *Inference to the Best Explanation*, 2 edn, New York: Routledge.

McDowell, J.: 1996, Two sorts of naturalism. Reprinted in McDowell (1998).

McDowell, J.: 1998, *Mind, Value, and Reality*, Cambridge: Harvard University Press.

McGee, V.: 2000, 'Everything', *in* G. Sher and R. Tieszen (eds), *Between Logic and Intuition: Essays in Honor of Charles Parsons*, New York: Cambridge University Press.

Millgram, E.: 2005, Practical reason and the structure of action, *in* E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/practical-reason-action/.

Moore, G. E.: 1903, *Principia Ethica*, New York: Cambridge University Press, 1998.

Morton, A.: 2003, *The Importance of Being Understood: Folk Psychology as Ethics*, New York: Routledge.

Nagel, T.: 1970, *The Possibility of Altruism*, New York: Oxford University Press.

Nagel, T.: 1989, *The View from Nowhere*, New York: Oxford University Press.

Nagel, T.: 1997, *The Last Word*, New York: Oxford University Press.

O'Neill, O.: 1989, *Constructions of Reason*, New York: Cambridge University Press.

Parfit, D.: 2011, *On What Matters*, New York: Oxford University Press.

Parsons, C.: 1977, What is the iterative conception of set? Reprinted in Parsons (1983).

Parsons, C.: 1983, *Mathematics in Philosophy*, Ithaca: Cornell University Press.

Pincock, C.: 2007, A role for mathematics in the physical sciences, *Noûs* **41**(2), 253–275.

Putnam, H.: 1962, The analytic and the synthetic. Reprinted in Putnam (1975).

Putnam, H.: 1975, *Mind, Language, and Reality*, New York: Cambridge University Press.

Putnam, H.: 2002, *The Collapse of the Fact/Value Dichotomy*, Cambridge: Harvard University Press.

Quine, W. V.: 1951, Two dogmas of empiricism. Reprinted in Quine (1953).

Quine, W. V.: 1953, *From a Logical Point of View*, Cambridge: Harvard University Press.

Quine, W. V.: 1960, *Word and Object*, Cambridge: MIT Press.

Quine, W. V.: 1981, What price bivalence? Reprinted in Quine (1986).

Quine, W. V.: 1986, *Theories and Things*, Cambridge: Harvard University Press.

Quine, W. V.: 1990, *Pursuit of Truth*, Cambridge: Harvard University Press.

Railton, P.: 1997, On the hypothetical and non-hypothetical in reasoning about belief and action, *in* G. Cullity and B. Gaut (eds), *Ethics and Practical Reason*, New York: Oxford University Press, pp. 53–79.

Rawls, J.: 1971/1996b, *A Theory of Justice*, Cambridge: Harvard University Press.

Rawls, J.: 1996a, *Political Liberalism*, revised edn, New York: Columbia University Press.

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

Reath, A.: 1997, Legislating for a realm of ends: The social dimension of autonomy. Reprinted in Reath (2006).

Reath, A.: 2006, *Agency and Autonomy in Kant's Moral Theory*, New York: Oxford University Press.

Reichenbach, H.: 1949, *The Theory of Probability*, 2 edn, Berkeley: University of California Press.

Rosati, C.: 2003, Agency and the open question argument, *Ethics* **113**(4), 490–527.

Rosen, G.: 1994, Objectivity and modern idealism: What is the question?, *in* M. Michaelis and J. O'Leary-Hawthorne (eds), *Philosophy in Mind*, Dordrecht: Kluwer.

Rosen, G.: 2010, Metaphysical dependence: grounding and reduction, *in* B. Hale and A. Hoffman (eds), *Modality: Metaphysics, Logic, and Epistemology*, New York: Oxford University Press.

Russell, B.: 1906, On some difficulties concerning transfinite numbers and other types, *Essays in Analysis*, London: G. Braziller, 1973. Originally delivered to the London Mathematical Society.

Scanlon, T. M.: 1998, *What We Owe to Each Other*, Cambridge: Harvard University Press.

Scanlon, T. M.: 2009, *Being Realistic about Reasons*. John Locke Lectures, Oxford University. Audio available at http://www.philosophy.ox.ac.uk/lectures/.

Schapiro, T.: 2001, Three conceptions of action in moral theory, *Noûs* **35**(1), 93–117.

Schechter, J. and Enoch, D.: 2008, How are basic belief-forming methods justified?, *Philosophy and Phenomenological Research* **76**(3), 547–579.

Setiya, K.: 2007, *Reasons without Rationalism*, Princeton: Princeton University Press.

Shafer-Landau, R.: 2005, *Moral Realism: A Defense*, New York: Oxford University Press.

Sluiter, I.: 2005, Communicating cynicism: Diogenes' gangsta rap, *in* D. Frede and B. Inwood (eds), *Language and Learning: Philosophy of Language in the Hellenistic Age*, New York: Cambridge University Press.

Smith, M.: 1994, *The Moral Problem*, Cambridge: Blackwell.

Smith, M.: 2009, The explanatory role of being rational, *in* D. Sobel and S. Wall (eds), *Reasons for Action*, New York: Oxford University Press.

Sosa, E.: 2009, *Reflective Knowledge*, New York: Oxford University Press.

Strawson, P. F.: 1962, Freedom and resentment, *Freedom and Resentment and Other Essays*, London: Methuen, 1974.

Street, S.: 2008, Constructivism about reasons, *in* R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Volume 3*, New York: Oxford University Press.

Sturgeon, N.: 1988, Moral explanations, *in* G. Sayre-McCord (ed.), *Essays on Moral Realism*, Ithaca: Cornell University Press.

Taylor, C.: 1971, Interpretation and the sciences of man. Reprinted in Taylor (1985).

Taylor, C.: 1985, *Philosophy and the Human Sciences*, New York: Cambridge University Press.

Toulmin, S.: 1972, *Human Understanding*, Princeton: Princeton University Press.

Velleman, J. D.: 1992, The guise of the good. Reprinted in Velleman (2000).

Velleman, J. D.: 2000, *The Possibility of Practical Reason*, New York: Oxford University Press.

Velleman, J. D.: 2006, *Self to Self*, New York: Cambridge University Press.

Velleman, J. D.: 2009, *How We Get Along*, New York: Cambridge University Press.

Wallace, R. J.: MS, Constructivism about normativity: some pitfalls. http://philosophy.berkeley.edu/file/495/Constructivism-Sheffield.pdf.

Weyl, H.: 1952, *Symmetry*, Princeton: Princeton University Press.

Wiggins, D.: 1976, Truth, invention, and the meaning of life, *Needs, Values, Truth*, New York: Oxford University Press, 1992.

Wigner, E.: 1967, *Symmetries and Reflections*, Bloomington: Indiana University Press.

Williams, B.: 1978/2005, *Descartes: The Project of Pure Enquiry*, New York: Routledge.

Williams, B.: 1985, *Ethics and the Limits of Philosophy*, Cambridge: Harvard University Press.

Wilson, E. O.: 1998, *Consilience*, New York: Alfred A. Knopf.

Wittgenstein, L.: 1969, *On Certainty*, Malden: Basil Blackwell.

Wood, A.: 1999, *Kant's Ethical Thought*, New York: Cambridge University Press.

Wood, J.: 2002, Mixed feelings. Reprinted in Wood (2004).

Wood, J.: 2004, *The Irresponsible Self: On Laughter and the Novel*, New York: Farrar, Straus and Giroux.

Woodward, J.: 2000, Explanation and invariance in the special sciences, *British Journal for the Philosophy of Science* **51**(2), 197–254.

Wright, C.: 1992a, The euthryphro contrast. In Wright (1992b).

Wright, C.: 1992b, *Truth and Objectivity*, Cambridge: Harvard University Press.

Wright, C.: 2004, On epistemic entitlement: warrant for nothing (and foundations for free)?, *Proceedings of the Aristotelian Society Supplementary Volume* **78**, 167–212.