2
Laws of Nature, Laws of Freedom, and
the Social Construction of Normativity

Kenneth Walden

In this essay I develop a theory of categorical normativity, of those principles that have authority over us regardless of our ends and interests. I argue that there is an intimate connection between these norms and the conditions of agency. In this respect, my account is a version of constitutivism. My constitutivism is unique in a few respects. First, it is naturalistic: agency is an emergent property, like the properties of biology and economics. Second, it is social: agency is something constructed by the complex interactions of agents. And third, it supports the categoricity of particular moral norms: adhering to one of Kant’s formulations of the Categorical Imperative, the Formula of the Realm of Ends, is a condition on agency.

It will help to have the structure of the essay laid out in advance. Section 1 is devoted to an argument that the conditions of agency do indeed produce categorical. In Section 2 I begin to sketch my account of agency. I criticize Kant, the grand pooh bah of constitutivists, for offering an account of agency that makes it mysterious how agents are capable of affecting the physical world. I then suggest that solving this problem while holding onto a conception of agency that can ground categorical norms is the chief obstacle to an interesting and plausible version of constitutivism. The essay as a whole can be seen as an attempt to overcome this obstacle. In Section 3 I turn to three well-known ideas about agency, and use them to suggest a conception that might fit the bill. Agency is an emergent property, and the study of agency is a special science. I further suggest that understanding the dynamics of this special science—what is required to explain a system as an agent—is a promising way to investigate the conditions of agency. In Section

1 The most sophisticated versions of constitutivism are due to Christine M. Korsgaard (1996) and (2009); and David Velleman (2000) and (2009). See also Connie Rosati (2003), Tamar Schapiro (2001), and Peter Railton (1997).
I argue that this special science has a distinctive methodology that sets it apart from other sciences: Davidson’s method of interpretation. From this discussion I conclude, in Section 5, with the first major thesis of the essay, that the laws of interpretation are conditions of agency and so we are categorically required to obey them. In Sections 6 and 7 I pause to field two objections. One charges that my account is mistaken because there is nothing particularly desirable about being intelligible. The other suggests that my view makes wrong action impossible. After dispatching with these objections I turn to the character of the categorical norms supported by my thesis. In Section 8 I consider but ultimately reject an argument that Kant’s Formula of Universal Law is such a norm. In Section 9 I turn to another formulation of the Categorical Imperative, the Formula of the Realm of Ends. Agency is special because it is the product of a complex interaction of actors interpreting each other and acting on the basis of these interpretations. For this reason it is a condition of agency that an actor participates in a negotiation with other actors over which principles are fit to be acted on. I then argue that this negotiation is identical to the contractualist procedure intimated in the Formula of the Realm of Ends, and so the latter is a categorical requirement. I close in Section 10 by drawing out the implications of this view and responding to an objection.

1. CONSTITUTIVISM AND THE SOURCES OF NORMATIVITY

Constitutivism is the thesis that the solution to the problem of categorical normativity—what we ought to do independent of our interests and aims—emerges from the very way that problem is introduced. The categorical norms of action are precisely 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.

For constitutivists, this is not an idle elegance. They see their view as 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, as Jones explains, he doesn’t care about \( F \)-style norms any more than \( n \) itself. So he remains unmoved.
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 merely hypothetical. If we want to show that some norm is categorical, we need to find some very special $F$ that Jones cannot parry away.

There are two models for this search. Both involve the questions of children. The first begins with the familiar thought that our explanations have to stop somewhere. A child asks, “why do cicadas have such an odd life cycle?” and we 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, and 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 bound to $n$, that this obligation is a bedrock law of nature.

The constitutivist looks to the 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.

The constitutivist models his response to Jones on this observation. Jones’s questions are part of the activity of practical reasoning, and, just like chess, this activity has certain conditions that must be met for someone to be engaged in it. So for Jones to ask, “why should I be interested in things that are $F$?” he must accept, at least implicitly, the conditions of practical reasoning. Among these conditions is that the person asking such a question be an agent. She is asking what she should do, so she must be the sort of creature who can do things, a creature whose deliberations affect the world through her will. Otherwise, this activity is not really practical reasoning, but idle speculation with a practical veneer. The constitutivist’s idea is to halt Jones’s regress in the conditions on generating that regress.

How does this show that the demands of agency automatically have authority over us? If the constitutivist is right, there should be something amiss with the question, “why should I be an agent rather than something similar but essentially different—a shmagent?” But to most ears this sounds

---

2 This question is at the heart of David Enoch’s (2006) challenge to constitutivism.
like a perfectly apt question. Even if the question is premised on one’s being an agent, it seems eminently fair to ask why one should carry on with that agency. Consider our child’s question, “why should I try to checkmate your king?” If we understand this question as a query about chess strategy, then it is confused. But we need not take it this way. We can understand it as a way of stepping outside of the game and asking a more general question, something like “why are the rules of chess what they are?” or “why should I play chess with you?” The concern for constitutivists is that the question “why should I be an agent?” is just as sensible as these questions, and for similar reasons.

But the question is not sensible because agency is special in a way chess is not. It is special first because of its connection to practical reason, and, second, because of a feature of practical reason I will call reflective closure. “Why should I play chess?” is not a how-to-play-chess question; it’s a which-game-to-play question. By contrast, the question “should I be an agent?” is itself a question about what to do, and so it is a question of practical reason. Wondering whether we should engage in practical reason is itself part of the activity of practical reasoning. For all activities X—or at least all the ones I can imagine—the question, “should I engage in X?” is not something we ask in the midst of X. Only when X is the broadest of all activities, practical reasoning, is this question part of X. Thus practical reason is reflectively closed in a way that no other activity is.

Agency is a condition on practical reasoning, and practical reasoning is reflectively closed, so agency is a condition on our asking the question “should I be an agent?” This fact does not show that the answer to the question, “should I be an agent?” is yes. (Saying that would be to give up on constitutivism, for it would concede that we need a deeper source of normativity to vindicate our constitutivism.) Instead, reflective closure gives us a way to articulate the defect in the question “why should I be an

3 Other defenders of constitutivism (notably Luca Ferrero (2009)) discuss a property of agency that is similar to reflective closure. They call it “inescapability.”

4 Enoch (2010, pp. 216 ff.) has a few replies to this line of thought. He protests that the inference from “O-ing is inescapable” to “I am justified in O-ing” is not valid. This is true enough, but irrelevant to my argument (and, if I read him correctly, Ferrero’s as well). Inescapability is not supposed to give us a further reason for being an agent. That would be contrary to the whole spirit of the program. It is supposed to show that there is something wrong with asking for such a reason. Enoch’s second reply concerns what he calls “dialectical inescapability”: the idea that there is nowhere for someone to stand and launch the shmagency challenge. Enoch responds that the shmagency challenge is not a dialectical challenge, but a question we put to ourselves. This is, once again, true, but beside the point, at least as far as my argument goes. My notion of reflective closure is not a dialectical notion, but a feature of agency we encounter from the first-person.
agent (instead of a shmagent)?” The question is defective because it has incompatible presuppositions. First, because it is a question asking for guidance about whether to be an agent, it presupposes that non-agency is a possible outcome of my deliberations. This follows from a general principle. The function of the questions “should I be an X?” and “what reason do I have to be an X?” is to initiate deliberations about whether to be an X, and it is obviously a presupposition of this kind of deliberation that I may end up a non-X. Thus, the question presupposes that I may answer it in a way that eventually leads to my non-agency. Second, because of the closure of practical reasoning, this question is itself a question of practical reasoning, and practical reasoning presupposes the agency of the person reasoning. The function of the question, “should I be an X?” is to initiate deliberations that will eventually issue in my deciding to be, or not to be, an X. But in order for me to do this particular thing, I must be able to do things in general. I must be an agent. So the content of this question, that it is asking about whether to be an agent, carries the presupposition that non-agency is a possible outcome of the deliberations guided by these questions, but the kind of question it is, that is, a practical question about what to do, presupposes that I will be answering it as an agent.5

With this we have two ideas for collapsing the hypothetical imperative. A certain kind of realist wants to halt Jones’s regress in fundamental norms modeled after the explanatory basic laws of nature. The constitutivist wants to halt the regress in the conditions of its generation: in the demands of agency. But the constitutivist 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 a practical matter 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 disposition. It would take a very queer entity to hold sway over my will just by dint of existing.

On the other hand, constitutivism locates the seat of normativity in the very
thing that makes normativity necessary. Our ability to act introduces the problem of how we ought to act, which is precisely 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 what to do than that the rules of chess hold sway over our deliberations about whether to castle.

2. THE QUEERNESS OF THE KANTIAN WILL

Every constitutivist has two tasks. She must explain why agency—of all things—can anchor a theory of normativity, and she must develop an account of agency from which a body of categorical norms can be derived—preferably interesting ones. In the previous section I took up the first of these tasks. The remainder of the essay is devoted to the second. My account of agency is developed in response to some misgivings about a conception of agency popular with other constitutivists.

Worries about queerness were a key part of the case for preferring the constitutivist’s solution to our regress to the realist’s. But if the constitutivist isn’t careful, she can end up with her own, homegrown queerness. In fact, the granddaddy of all constitutivisms, Kant’s, is guilty of just that. Kant’s argument that the moral law is constitutive of agency depends on the will’s total independence from the physical world. Here, for instance, is an argument from the second *Critique*:

Since the matter of a practical law . . . can never be given otherwise than 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.6

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.*7

---


And this gets us a hop, skip, and a jump away from the first formulation of the Categorical Imperative, the Formula of Universal Law. For Kant it is because the will is “independent” of the empirical world that it is directed by the pure “form” of lawfulness in the way the Formula of Universal Law requires. So it is the thoroughgoing independence of the will that gets Kant his constitutivism.

I don’t think we can live with this kind of chasm between will and world. In liberating the will from the empirical world, Kant also robs it of its characteristic ability, the ability to affect that empirical world. Now Kant would deny this. He thinks our decisions do affect the empirical world. But they do so through the inscrutable mechanism that Kant calls noumenal affection. Our will, alongside all other intelligences, configures the intelligible world, and this world, in turn, undergirds the empirical world in the inexplicable way that things in themselves undergird appearances. But this two-world account of how my decisions make a difference in the order of the physical world is just as queer as the realist’s postulate. Only here the queerness runs in the opposite direction. The realist posited a mysterious normative power that some objects hold over the will. 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.

There are, of course, other strains of constitutivism, but instead of painstakingly developing this objection against them one by one I want to 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. 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.

What we need, then, is a middle ground between a realism that places normativity too remote from the will and a Kantian constitutivism that places the will too distant from the world. This is the space I will try to occupy. I locate normativity in the requirements of agency, but see agency as a part of the physical world.

3. AGENCY AS AN EMERGENT PROPERTY

The key to this undertaking is finding a way for agency to be a part of the physical world without it dissolving into the background noise of particle physics. Toward that end, consider three now-familiar lessons about agency.
The first of these lessons comes from Harry Frankfurt. Imagine a man who has been surgically altered 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 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 agency should not be cashed out in terms of counterfactuals about what a person could and could not have done, but as a capacity for manifesting one’s preferences, values, and commitments. Capacities can be undercut of course: when our interloper interlopes, our man’s agency is lost. But just because this capacity can be taken way doesn’t mean it was never there.8

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 vowing 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 can of paint, 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 in his heart. What lies behind this dichotomy of reaction is the thought that one form of explanation is appropriate for inanimate objects, while another kind is appropriate for people.9

Marrying Strawson’s 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.10

The third lesson comes from Daniel Dennett. The special mode of explanation appropriate to agency is a kind of pattern recognition. We are confronted with a sea of behavior that might look like noise, but we can

8 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, for instance, to hold someone responsible.

9 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 up the reactive stance cannot mean treating someone like a black box. What’s distinctive of this stance is the kind of understanding we seek.

10 Note that this is just a material biconditional, and neither side is being analyzed in terms of the other.
make sense of it by recognizing 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 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 same goes for action. 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 particles. But if we look instead for patterns that emerge at a coarser grain of resolution, at beliefs, desires, intentions, reasons, and choices, then we may begin to predict and explain the behavior of these complex systems.11

These lessons offer two potentially fruitful ideas for constitutivists. The first idea is that agency is an emergent property, just as the social features of ant colonies are, and the study of agency is a distinctive kind of special science. This conception of agency fits well with our Goldilocks moral. 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 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.

The second fruitful idea amounts to a Copernican turn in our approach to constitutivism. Putting our three lessons together we find that 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.12 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 is, in effect, a prerequisite for being explicable as a free agent, and hence a prerequisite for agency itself. With this idea in mind, we can develop our constitutivism

11 Dennett (1991). There is a vast literature, much of it more rigorous, on what features of the world ground the use of higher-level sciences like these. I cite Dennett for the perspicuity of his simile.
12 That is to say: not merely what patterns will in fact lead us to ascribe agency, but which patterns ought to lead us to do so.
not simply by looking directly at agency, but by examining what it takes to understand a system as an agent. These requirements will then correspond to requirements of agency and, thereby, to categorical norms. From this we get our first thesis:

**Agency Emergence.** Agency is an emergent property that we ascribe to a complex system when it is appropriate to explain that system as an agent.

Going forward, I will investigate the conditions on this understanding as a way of coming to grips with agency itself.

### 4. REASON AND INTERPRETATION

What does it mean to understand a system as being an agent? What kind of explanation does this amount to?

There is a temptation to assimilate explanations of agency to some already established scientific field. We might say that understanding someone as an agent is just to understand her as being human in some important respect. This approach would subsume agential explanation to anthropology or human psychology. Neo-Aristotelians, like Philippa Foot, hold a view like this. They think that certain generalities about living creatures (what are sometimes called “Aristotelian categoricals”) ground evaluations of those creatures in special cases (roughly, when the generalities are teleological in character). “Owls see in the dark” is an Aristotelian categorical that grounds the evaluation that “an owl that cannot see in the dark is a bad owl.” Foot’s idea is that Aristotelian categoricals about humanity correspond to norms.

This essay is not a critique of Foot’s program, but there is one problem with her approach that is instructive for constitutivists. Aligning agency too closely with human nature, and thus agential explanation with anthropology or psychology, threatens to turn all of humanity’s warts into normative ideals. It is rare to find someone who doesn’t commit the conjunction fallacy or waver in the face of temptation. It seems fair to say that these features are part of human nature. They may even offer some survival advantage; if not to us, then to our lineage. Nonetheless, the conjunction fallacy is still a fallacy, wavering in the face of temptation is usually a mistake, and it would be ridiculous to call these things categorical requirements.

However the Aristotelian wants to handle this problem, it has obvious implications for the constitutivist. Her theory must be grounded in a conception of agency that is narrower and more refined than straightforward generalizations about human beings. The constitutivist cannot completely avoid this problem. Indeed, no theory built upon a deep connection between normativity and humanity can. We can substantially mitigate the problem by understanding agency as something that human beings are capable of without achieving by default. On the other hand though, we cannot just lay down a conception of agency by fiat, lest we end up reverse engineering something that validates the lion’s share of our common-sense judgments about morality and rationality. It would be very dubious indeed to stipulate that agency is the human tendency to uphold the commandments of the Hebrew Bible. The resolution to this dilemma is simple enough. The concern with reverse engineering turns on the question of whether there are phenomena distinctive enough to warrant the introduction of a special explanatory apparatus above and beyond what is offered by traditional psychology and anthropology: whether positing a special capacity of “agency” might allow us explain things that a more flat-footed accounting of human nature cannot. So are there?

There plainly are. Even better, there is a familiar mode of explanation naturally attached to them. The phenomenon is our ability to control our behavior through reflective reasoning, and the distinctive mode of explanation that accompanies it is the method of interpretation. I will say a word about each in turn. It may be altogether normal to be tripped up by the conjunction fallacy or waver in our resolutions, but we are not doomed to be in the thrall of these vices. Nor must we see them as irredeemable parts of our nature. We can view them as defects, and we can call them as much when we see them in other people because we have the ability to transcend these tendencies through reason: to detach ourselves from our instincts, hold them up to scrutiny, and, if appropriate, renounce them. Thus reason allows us to overcome the part of our nature that leads us into these mistakes, and to expect the same elevation in other people.14

Now, we must distinguish between reason itself and the ability of human beings to regulate their behavior through reason. The conception of agency I am interested in is the latter. This capacity requires a great deal more than the mere ability to partake in reflective reasoning. To see why, we need look no further than akrasia. I am inclined to eat this cake, but I stand back from that inclination and judge that, all things considered, I ought not eat the cake.

14 This conception of reason is a very thin one: compare it to the conception of reason that sees it as a faculty for detecting “intellectual seemings” and informs us about the “rational structure of the universe.” It is also Kant’s. See A795/B823.
cake; and yet despite this judgment, my inclination hijacks my motivational system and I eat it anyway. Here my power of reason is fully functional, but my judgments of reason do not regulate my behavior.\(^{15}\)

What does agency so understood have to do with the method of interpretation? First things first: what is the method of interpretation? William Child has described it limpidly. “When we interpret someone,” Child explains, “we aim to make sense of her by attributing beliefs, desires, intentions, emotions, and other propositional 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.”\(^{16}\) The key word in all this is “rational.” We can imagine an approach to explaining behavior that trades in the same attitudes as the method of interpretation—belief, desire, and so on—but posits these attitudes as brute causes of behavior, as causes independent of any rationalizing constraint on our explanation. A belief and desire jointly cause one’s behavior in just the way that an acid and a base cause the production of salt.\(^{17}\) What distinguishes the method of interpretation from this pattern of explanation is the idea that interpretation aims, when possible, to offer a rationalizing explanations. For an explanation to be an interpretation, it cannot just offer any attitudes that tend to cause the explanandum. It must make sense of the action as a rational thing to do.

This is where the aptness of the method of interpretation to the phenomenon of agency-produced behavior emerges. The method of interpretation’s notion of rationality should be understood in terms of the notion of reason attached to agency. An action is rational just in case it lives up to the scrutiny levied by the reflective power of reason. The method of interpretation is therefore the method of explanation appropriate to behavior that has been successfully regulated by reason, which is just to say that interpretation is the method appropriate to explaining the behavior of agents.

Child’s gloss may make the method of interpretation sound more fixed and immutable than it needs to be. A sketchier, but more accurate picture of the method is this. We observe certain phenomena: human behavior guided, in one way or another, by reflective reasoning. Many different

\(^{15}\) Compare Michael Bratman’s (2007) theory of agency.

\(^{16}\) 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).

\(^{17}\) We might think that it is part of our concepts belief and desire that they are ascribed as part of a rationalizing explanation. Most interpretationists believe as much, so what we are imagining here are different states that go by the name “belief” and “desire.”
theories of human behavior are consistent with these phenomena, and so to overcome this underdetermination we have to introduce constraints on our explanations. We introduce new constraints until we are able to winnow down the class of explanatorily adequate theories to a single account. (Perhaps we’ll never get a single account, but we can try to overcome the underdetermination as much as possible.) 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.” So it is not gospel that our interpretations 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. These are good candidates for the constraints because they give us good explanations, but nothing more. 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. We begin with an inkling of both. Agency is something like our ability to regulate our behavior through the reflective power of reason, and interpretation is something close to our ordinary ways of understanding our friends and neighbors. The two then evolve symbiotically. (This is not an unusual arrangement either. We see the same co-evolution of explanans and explanandum in other special sciences, for example, in the long and tortuous relationship between genetics and inheritance.)

The particulars of the method of interpretation will matter very little going forward. What matters is the boundary between interpretation and other explanatory strategies. Sometimes interpreting a person 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 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

---

18 Lewis (1974), p. 343. I actually disagree with Lewis about our ability to reach a single theory, but this departure is inconsequential for our purposes.
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, or delirious with joy. Not all mechanical explanations are clunky and hydraulic. Mechanical explanations include anything that attempts to assimilate a person’s behavior to an 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 the instinctive grooming habits of chimpanzees, the 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 paradigms of a diverse array of explanatory rubrics, each associated with a different science. Each of these sciences has the potential to be adapted to offering an explanation of certain kinds of human behavior, sometimes even a better explanation than interpretation.19

A summary is in order. In the first section I presented a general argument for constitutivism. The conditions of agency correspond to categorical norms because agency is a condition of practical reasoning, and normative discourse only makes sense within the activity of practical reasoning. In making this argument I offered no substantive conception of what practical reasoning is, and thus no definite conception of agency. We now have such a conception: reason as a power of critical reflection.20 With this conception of reason comes a substantive conception of agency: the capacity to control our behavior through reason so understood. Agency so conceived is the subject matter of a special science, and the methodology particularly suited to this special science is the method of interpretation.

19 I do not see interpretation as radically different from other forms of explanation. On this point I depart from defenders of interpretation, like Charles Taylor (1971), who think that interpretation is an alternative to explanations that invoke traditional notions of causation, regularity, and prediction. I doubt that something can count as explanation if it eschews these categories too aggressively. What distinguishes interpretation on my view is the rationalizing constraint on explanation, not a total abandonment of explanation.

20 Were we differently inclined we could pursue a constitutivist project that takes this thin conception of reason as its target. This is one way of understanding Christine Korsgaard’s project in her (1996). I am pessimistic that we can get much normative content out of a constitutivism built upon such a thin reed.
5. SUBSUMPTION UNDER LAWS

The previous section presented two ideas. First, a subject is an agent if and only if it is appropriate to explain his behavior in a very particular way. Second, this distinctive method of explanation is interpretation. Putting these two ideas together yields our first lemma:

Lemma 1. The extent to which a subject is an 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.

Consider the following three examples of special science laws. 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. 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. 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, something is lost in translation.

This is a corollary of the much more general thought that all explanation is a matter of subsuming some 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.21 The letter of the covering law model founders on familiar counterexamples, but its spirit 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. If we understand the subsumption

---

21 Hempel and Oppenheim (1948). For a version of this idea shorn of Hempel’s logicism see Michael Friedman (1974).
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 uncontroversial.\textsuperscript{22}

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 special science, then it too must have distinctive laws.\textsuperscript{23} These laws will be \textit{ceteris paribus} ones, of course, and a good deal rougher than the laws of physics. But the basic point survives this proviso. 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:

\textit{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 together our two lemmas we get:

\textit{Agency-Interpretation Thesis.} 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.\textsuperscript{24}

This thesis is little more than an explication of the lessons outlined above. Agency is intimately connected to how we react to and explain a subject’s behavior, and what this ultimately amounts to is the relationship enshrined in the Agency-Interpretation Thesis. When we combine this thesis with constitutivism we get:

\textit{Generic Interpretative Constitutivism.} We are categorically required to adhere to the laws of interpretation, whatever they may end up being.

We become aware of the normative force of the laws of interpretation in a moment of self-awareness. Our interpretations of other people contribute to a theory of the laws of human agency. But we \textit{qua} actors are subject to

\textsuperscript{22} James Woodward (2000) argues that explanation in the special sciences involves invariance relations and \textit{not} laws. I think the very weak understanding of laws I am offering here manages to accommodate even his view, which is the most extremely anomic position I know.

\textsuperscript{23} It is worth noting that the invocation of laws in Generic Interpretative Constitutivism makes it incompatible with many forms of particularism. This is a congenial consequence for me, since I am not a particularist. But were someone so inclined, she could remove the nomic assumptions I make here and get a corresponding version of the thesis that is friendly to particularists.

\textsuperscript{24} This thesis can be understood as something akin to what some philosophers of science (e.g. Lange 2009, pp. 104–10) call “meta-laws.” These principles do not themselves explain. Instead they state principles about what an explanation must look like to be successful.
the laws that we qua interpreters discover: we are one of the creatures whose behavior is governed by these laws.

Some product differentiation may make Generic Interpretative Constitutivism clearer, so I close this section with that. Recent years have seen two other writers plumb the relationship between interpretability and ethics. The first is David Velleman, who holds that the constitutive aim of action is self-understanding. 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. Velleman’s thesis is about the telos of action, mine just about the conditions of a thing being an 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. This kind of claim about taxonomy is very much at home in science. It is quite another to say that having such and such an aim is constitutive of being a chromosome or a galaxy or an agent, and it is much harder to find evidence of this kind of teleology in science.

A second difference is that I do not begin with a detailed a priori account of the requirements of action. I will eventually go on to show how we can make out the contours of these requirements by reflecting on the structure of explanation, but for now all I do is show why there are such requirements and what kind of things they might be. Velleman, by contrast, begins 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 by analogies between action and belief. He thinks belief has an obvious and a priori graspable aim—truth—and is interested in finding the analogous aim for action. Velleman’s claim about belief strikes me as too simplistic. What is good in the way of belief is a very subtle question bound up in issues about the relationship between belief and understanding. But whatever we think about belief, the idea of a single, constitutive aim is all the more suspect in the case of action. Sartre surely had it right that human beings are not like paperknives, and we are not created for a particular purpose. The whole game of constitutivism is to find some account of the nature of action that can produce informative categorical norms while still holding onto this basic point.

Finally, Velleman is interested in our interpretability only insofar as it enhances our self-understanding. Being interpretable to others, he says,

25 See the aforementioned (2009), as well as his (2000) and (2006).
helps us engage in repeatable scenarios that make our behavior more intelligible to ourselves, and it is worthwhile for just that reason. In a similar vein, Adam Morton emphasizes the importance of interpretation to coordination problems between agents in which the best solution for all agents is only attainable if they can understand each other.\textsuperscript{26} I see a much more direct connection between intelligibility (that is, being interpretable) and agency. Intelligibility is not merely a means to the \textit{real} goals of practical reason (self-understanding, utility maximization, or what have you). Interpretation tracks what it takes to be an agent, and so its laws are fundamental norms of practical reason.

6. FIRST INTERLUDE: THE PROBLEM OF CALCULATED INSCRUTABILITY

Let me pause to field two instructive objections. What do I have to say about all the people who are not only difficult to interpret, but actually cultivate inscrutability to great success?

The songstress Lady Gaga once wore a dress made of meat.\textsuperscript{27} 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 Dali 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, “Dali is immortal and will not die.”

Gaga, Dylan, and Dali 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 I should say to them, “insofar as you want to do anything at all, from wearing a suit made of meat to stealing pens, you have to be an agent, so you really must obey these rules.” This appeal seems feeble in the face of the fact that Gaga, Dylan, and Dali \textit{are} successful very much \textit{because} they violate the demands of interpretability.

But these examples can be misleading. Our three oddballs behave strangely enough to render the standard arsenal of heuristics we use to

\textsuperscript{26} Morton (2003).
\textsuperscript{27} Thanks to the many questioners in Madison who put the problem of Lady Gaga to me so forcefully.
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 falls short when we encounter Gaga and her meat dress. And this failure of our superficial methods of interpretation leaves us flummoxed. But this initial bemusement need not augur a complete failure of interpretation. We have enough data about Gaga to retreat and give an explanation of her behavior that does not rely on heuristics about the relationship between season and fabric. We can give an explanation that starts from a more basic place. One obvious interpretation sees Gaga with a jaundiced eye. She wants fame and fortune, and she has decided that frustrating our day-to-day expectations is a good way to get them. 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. A more subtle interpretation sees Gaga’s behavior as an artistic challenge to the conventions that make our usual heuristics effective. As Marcel Duchamp put a urinal in an art museum to upset our expectations about what we might see there, so Gaga puts on a meat dress to upset our expectations about what a popular singer wears. Her goal is didactic: to shake up the public mind and allow us to see our preconceived notions as optional. (This is an important species of action that I will return to later.) At the end of the day, these eccentricities may yet prove incompatible with full-fledged agency. But the important point is that despite its superficial intractability, the method of interpretation offers the resources to make a start on understanding Gaga’s 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 happiness, 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 actual wit, the philosopher who jokes, “I do everything I can to be difficult to understand”\(^\text{28}\), or one’s third hour in a Dadaism exhibition.

A good example of this predicament can be found in the tricky business of creating a literary 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 tedious inscrutability.

\(^{28}\) Nietzsche, *Beyond Good and Evil*, section 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.29

While deliberate pseudo-inscrutability may be easy, genuine inscrutability is hard to pull off. Gaga, Dylan, Dali, and even Zeno 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.

7. SECOND INTERLUDE: DEGREES OF AGENCY AND THE PROBLEM OF BAD ACTION

Even if I am right in all I have said about calculated inscrutability, a deeper problem seems to remain, one that afflicts all forms of constitutivism. If following certain norms is a requirement of acting, then it is impossible for a person to act in violation of those norms. But what sense does that make?

There are a few objections in this question. The most menacing is this. Norms are supposed to guide our actions, and they do this by partitioning potential actions into classes of right and wrong, obligatory, permissible, and forbidden, or something similar. But according to the constitutivist, wrong actions are not actions at all. This means, among other things, that categorical norms are vacuous.

The solution to this problem lies in the gradedness of agency, something my theory provides for. My characterization of interpretation and mechanical explanation makes the line between them blurry. It is often unclear, for instance, whether a violent tantrum issues from rage overcoming a person, or from a decision to express that rage through an outburst. In these cases we hedge. We offer a story that has aspects of both mechanical explanation and interpretation, that sees the behavior as partially intelligible and partially

mechanical. This case and many like it make it impossible to mark off where interpretation ends and mechanical explanation begins with any precision. This is not just an epistemic point. It is not that all behavior is determinately agency or non-agency, and we just have a hard time telling one from the other. There are genuine middle-ground cases. (Most people are probably middle-ground cases.) This means that interpretability and agency admit of degrees. The question is not whether we interpret a system, but to what extent our best explanation is an interpretation; not whether a subject is an agent, but how much of an agent she is.

This fact about agency gives us a simple answer to the concern about the vacuousness of norms. A person’s behavior must be interpreted holistically, not one act at a time. We make sense of a person’s behavior over long stretches of time, and we attempt to come to grips with it as a unified whole. This means that discrete bits of behavior that are incompatible with the demands of agency do not vitiate a person’s agency at the moment they are performed, and so they are not automatically non-acts. They merely diminish the agency displayed over an extended episode of acting (the execution of a plan, a conversation). Indeed, the questionable acts themselves are rendered partially intelligible by their place in a more-or-less intelligible body of behavior. Assume, for instance, that a minimal capacity to adhere to one’s resolutions is a constitutive requirement of agency. So when LeNoir abandons his resolution to go straight home from work by stopping by the chocolatier, that makes him less intelligible as an agent. But if LeNoir strongly desires chocolate, his going into the store is partially intelligible—more intelligible, for instance, than it would have been had he hated chocolate. Even if an act cannot be integrated into a full and comprehensive interpretation, its place in a partially interpretative explanation, like the one we can give of LeNoir’s walk home, renders it partially intelligible. This graded view of action and agency gives us a corresponding account of how norms guide us. Norms connected to the constitutive requirements of agency do not partition potential behavior into actions and non-actions, but into things more and less consistent with the demands of agency. Of course, practical reason is conditioned on full-fledged agency, not half-bore agency or some minimal threshold of agency. So it is the conditions of full-fledged agency that our constitutivism entails are categorical requirements.

The second way our vague worry about the possibility of bad action might be turned into an objection concerns our secondary moral concepts. If the violation of a norm is something less than an action, then what sense is there in holding the violator to account or parceling out blame? My answer here apes the one compatibilists give to similar worries. We should not peg our concepts of responsibility, accountability, and blame to the
libertarian’s notion of freedom, nor should we peg them to full-fledged agency. Whether to hold someone responsible and blame him depends on how much of an agent he is, what particular deficiencies of agency he suffers from, and how holding him responsible fits into our legal and social practices of blame and punishment. Therefore, the fact that wrong actions diminish agency is no obstacle to our holding people responsible for them.

The final way to press our worry about bad action is to point to some act that is in violation of a norm that a given constitutivist says is constitutive of agency, and then insist that this is a perfectly good action. Of course, saying this is just to disagree that a particular principle is indeed a requirement of agency. But since I have not yet articulated any such principles, there is not much to say at the moment. But I will offer one admonishment. There is a tendency to conflate instrumental intelligibility with intelligibility tout court. Doing this makes it appear that intelligibility is only a matter of fit between ends and actions. Thus if we can make sense of the villain’s dastardly deeds given his villainous ends, then we are inclined to think his behavior is wholly intelligible. This is a legitimate theory of intelligibility, but it is not the only one, and it is certainly not something we should accept without argument. Consider again Monsieur LeNoir and this time his friend Monsieur LeBlanc. LeNoir has a long history of chocolate fixation: he bakes a chocolate cake twice weekly, owns a tasting set certified by the Ministère de Chocolat, reads treatises on cacao density, and regularly proclaims himself a missionary for the chocolate religion. LeBlanc by contrast finds the taste and texture of chocolate quite unappealing. Yet, as it happens, both of them have a desire for chocolate as they walk by the chocolatier. Must we really say that LeNoir and LeBlanc’s respective detours into the chocolatier are equally intelligible? Clearly not: they have the same desire for chocolate, but this desire is much more intelligible in one of them. This appears, then, to be a case in which intelligibility transcends instrumental intelligibility.

With these objections defanged, my case for Generic Interpretative Constitutivism is complete. In the next sections I turn to the task of putting meat on the thesis’ bones. Generic Interpretative Constitutivism offers little normative content. It doesn’t give us precise advice about what we ought to do; in particular, it is hard to see any progress toward the conclusion that moral requirements are categorical. I consider two approaches to drawing more content out of Generic Interpretative Constitutivism. First I show how the thesis could be made to entail a version of Kant’s Formula of the Universal Law, and how it offers some insights into the application of that notorious requirement. I ultimately reject this approach because it relies on extremely strong and unwarranted restrictions on how we are to explain behavior.
8. A FALSE STEP: UNIVERSAL LAW

CONSTITUTIVISM

One way to supplement and give content to Generic Interpretative Constitutivism is to assume that there is but one kind of practical principle. Kant pursues this kind of strategy. The beginning of *Groundwork* II is occupied with arguments against naturalistic and sentimentalist theories of the moral law. Kant argues that these views do not do justice to the universality of moral requirements. Everyone is bound by moral requirements, not just those with sentiments that incline them to moral feeling. This critique is central to Kant’s positive characterization of the moral law. Not only is universality a feature of the moral law, the moral law is the requirement that our practical principles be capable of being “universal laws of nature.”

Never mind whether this move is valid. Kant has a bigger problem. Universalizability by itself is no constraint at all, for there will always be some way of characterizing the practical principle behind an action that is capable of being a universal law. If we specify the conditions of an act so exhaustively that the act is utterly unique, then the requirement is satisfied trivially. The requirement stops being trivial only when joined with constraints on which principles are candidates for universalizability. This is where Kant makes a momentous assumption. He effectively assumes that all action can be cashed out in terms of one kind of principle, and that this is the only kind of principle we test for universalizability. The species of principle is the *maxim*, an attitude that directs the will by connecting means to end, one which we could articulate as “I will φ in order that ρ.” Joining universalizability to Kant’s Maxim Assumption, as I will call it, yields an important special case of the first formulation of the Categorical Imperative:

*Formula of Universal Law (Perfect Duties).* Act only on a maxim that you can conceive of being a universal law.30

This requirement can be given a constitutivist foundation if we join my constitutivism and Kant’s Maxim Assumption. I will ultimately reject Kant’s Maxim Assumption, but developing this kind of defense of the Formula of Universal Law does offer some insights into what Kant’s thinking behind the Formula of Universal Law is.

30 The *Formula of Universal Law sans phrase* requires not only that one be able to conceive of a maxim being a universal law, but that one be able to will it as such. Willing is a strictly stronger requirement than conceiving. See 4:421–4. Maxims that can be conceived of as universal laws but not so willed correspond to merely imperfect duties.
According to my view, our ability to explain an action depends on the actor’s practical principle being modally robust. Without this robustness, the connection between principle and action would be too accidental to be explanatory. Thus, we explain actions using the laws of agency—the actual laws, not just principles that could be laws—because those are the principles capable of furnishing genuine explanations. Again, by itself this requirement is pretty innocuous because we have no idea what the laws of agency might look like. But if we conjoin this feature of my constitutivism with Kant’s Maxim Assumption and assume that the only laws of agency are maxim-based laws, then we get a powerful conclusion. If a person violates the Formula of Universal Law, then his action could not instantiate a maxim-based law about how people act. But those laws, per Kant’s Maxim Assumption, are the ones we must use to explain an agent’s behavior. So such a person is uninterpretable. This yields a constitutivist foundation for the Perfect Duties version of the Formula of Universal Law.

Universal Law Constitutivism. Adherence to the Perfect Duties version of the Formula of Universal Law is a requirement on agency, and therefore a categorical requirement.

Universal Law Constitutivism requires that the connection between adopting a maxim and performing an action be lawful. But what do I mean by lawful? Suppose we adopt a counterfactual analysis of laws and a standard story about counterfactuals: “if it were the case that A, then it would be the case that 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. If we assume both of these things, then a maxim like “I will bake a cake in order that I have something to eat” instantiates a law if and only if everyone in nearby possible worlds who wants something to eat and chooses baking a cake as a means to that end goes on to bake a cake. The maxim-adopting worlds must be a subset of the cake-baking worlds. If this condition is not met, then the principle will not be a law, and our subject will not be interpretable by the ascription of a maxim.

The counterfactual account analysis is one way of making precise the idea that laws must represent genuine regularities and not just accidental generalities. And this is what gives Universal Law Constitutivism its teeth. Suppose, fantastically, that it is a law of nature 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. This means that there is a world nearby in which an agent is in a yellow room, adopts a

31 Lewis (1973).
maxim to bake a cake, but, because of this strange law, cannot effect the baking of the cake. This world is a witness to the falsity of the would-be law connecting the adoption of cake-baking maxims and cake-baking. We cannot appeal to such a maxim in an interpretation of our agent’s actions because its connection to our explanandum is too serendipitous to support an explanation. This conclusion may sound surprising, but it stands in close analogy to familiar cases. Suppose we have a machine that is so poorly constructed that it very seldom works. If we witness one of the rare occasions on which it does work, we cannot explain this success just by saying “that’s the way the machine was designed to work.” That’s not an adequate explanation because the same facts about design hold in all the nearby possible worlds where the machine didn’t work. We either need to find some more specific explanation that doesn’t have this problem (the machine only works in exceptionally high humidity) or concede that it worked out of dumb luck and we have no good explanation for it.

This analysis comes with a few serious caveats. First, the requirement that an undermining possible world be nearby 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. Second, we are dealing with *ceteris paribus* laws, so even within the truly nearby possible worlds we must tolerate some exceptions. Third, failing to reach a fully robust law only diminishes the power of explanation. Saying that our machine did something because it was designed to is some explanation, even if not an altogether complete one. These three caveats reduce the range of things Universal Law Constitutivism can forbid and require.

Universal Law Constitutivism offers a new way of understanding Kant’s “conceivability” test for non-universalizable maxims. Suppose Brown resolves, “I will make a false promise to get 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 promise at all, let alone a false one. 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 law connecting Brown’s maxim adoption to his action, then it falsifies that law. In this world adopting the maxim “I will make a false promise to get some ready cash” does not lead to the action of making a false promise because there is no such thing as promising. Thus because there is a world where adopting the fast cash
maxim does not lead to promising, the connection between that maxim and promising is not counterfactually robust enough to be a law. And this means we cannot interpret Brown’s action by citing this maxim.

This is how Kant’s conceivability test appears through the prism of Universal Law Constitutivism. I actually doubt that the test shows what Kant wants it to, that making false promises is categorically forbidden. These doubts arise from the caveats I mentioned a moment ago. The worlds in which the institution of promising breaks down are actually rather remote in logical space, so it is not clear that any of them are actually relevant to the putative maxim-action law. And if any of them are, it seems fair to say that they represent a few exceptions to a perfectly good ceteris paribus law.

My rendering of the conceivability test also appears to be rather different from Kant’s own. Kant looks at just one world, “the world of the universalized maxim,” in which everyone who possesses a given maxim’s end takes up that maxim. A maxim passes this test if and only if this world is conceivable. My test is concerned with nearby possible worlds in general, and with the world of the universalized maxim only insofar as it is one of these. Kant evidently sees the idea that we must be able to conceive of our maxims as “laws of nature” differently than I do. I see it as a matter of regularity, and hence modal robustness, whereas Kant connects it to the possibility of everyone acting in the same way. These two interpretations overlap, but they are obviously different. For my part, I can see quite clearly why we should be interested in regularity—for all the reasons I have given here—but fail to understand what is so important about the world of the universalized maxim.

But this question is moot, because we ought to reject Kant’s Maxim Assumption and with it the version of Universal Law Constitutivism explored in this section. There are two reasons for rejecting Kant’s assumption. The first is that it is false. The maxim is too narrow an instrument of interpretation to carry the burden of all interpretation. It is a very flexible practical principle, but we act in too many different ways for us to determine a priori whether all of them can be shoehorned into a single species of principle. The slightest acquaintance with our practical psychology shows it to be too complex, too balkanized, and too multifarious for Kant’s Maxim Assumption to be correct.

The second reason for rejecting Kant’s assumption is that most of the infamous problems with the Formula of Universal Law can be laid at the doorstep of this assumption. All the familiar false positives (“In order to avoid the crowds, I will play tennis on Sunday morning”) and false negatives (“In order to get some ready cash while still being interpretable, I will make a false promise if and only if my promise will be accepted”) that plague the Formula of Universal Law follow from this assumption.
Without the assumption, all that’s left of the Formula of Universal Law is a generic universalizability requirement, and that is completely innocuous.

9. REALM OF ENDS CONSTITUTIVISM

Without Kant’s Maxim Assumption, we are tossed back to the thought that one is required to follow the laws of interpretation—whatever those might be. But our hopes of getting a more informative form of constitutivism aren’t dashed quite yet. We may not come up with a principle requiring and forbidding highly specific lists of actions, but we can get the next best thing. We can derive a rough schema of categorical norms by showing that the demands of agency lead us into a popular legislation of practical principles. In other words, we can show that agency requires us to abide by whatever emerges from some contractualist procedure. Hobbes thinks that we are bound to the outcome of one such procedure because it is collectively rational, and Rawls thinks we are bound to the outcome of another because it enshrines our existing ideas about what is reasonable. The view I propose is that we are obligated to participate in a similar procedure because it is a requirement of being an agent.

Why are we obligated to participate in such a procedure? Earlier I suggested that a special self-awareness allowed us to recognize the truth of Generic Interpretative Constitutivism. We become aware that the I who explains agency and the I who act are one and the same, and, consequently, that what I am explaining are the laws of agency for me. Now I suggest that we recognize the authority of a primitive contractualism from the further awareness that everyone else is engaged in precisely the same endeavor: interpreting others’ behavior, systematizing those interpretations into rudimentary laws, and acting in accord with the laws so determined. My efforts are inextricably bound up with theirs because we jointly create the behavior present in the world—the very phenomenon that the laws of agency are laws of. Therefore, each person’s success in this process crucially depends on coordination with everyone else.

Call the sum of all this activity the Mutual Interpretation Process. I shall argue that this procedure is identical to the contractualist procedure found in the final formulation of the Categorical Imperative, the Formula of the Realm of Ends. This Formula requires that people “act in accordance with the maxims of a member giving universal laws for a merely possible realm of ends.”

32 Forget the restriction to maxims and instead take the Formula...
to apply to practical principles of all stripes. The Formula of the Realm of Ends imagines a legislature in which every member of a community puts forward practical principles as potential laws for every member of that community, and it requires that we act only on principles that would be universally ratified by such a legislature. The Formula of the Realm of Ends is an ur-contractualism: it lacks the idealizations of Rawls’s contractualism and the rationalist foundations of Scanlon’s. It will therefore not offer immediate and specific advice that one must do this thing and avoid that. Instead, it engenders two open-ended duties: to participate in a procedure of reciprocal legislation that will eventually produce this kind of guidance, and to adhere to the edicts of this legislature. Eventually, Kant promises, this procedure will bring us to an idea of the Realm of Ends itself, and that will offer “a complete determination of all maxims.” What I hope to show in this section is that the requirements of agency in a social medium coincide with the Formula of the Realm of Ends.

I proceed by adumbrating the Mutual Interpretation Process step by step and then comparing it to the Realm of Ends procedure. The first step is simple enough.

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

The explanandum of our theory of agency, and so the primary constraint on our conception of the laws of agency, is constituted by people’s actions. What behavior exists to be explained by a theory of agency is determined by what people do, and so the phenomenon we are trying to explain when we posit laws of agency—the phenomenon they are laws of—is partially man-made. (“Partially” man-made because other laws of nature impose constraints on what kinds of actions are possible: there is no such action as spontaneously turning into a pumpkin.)

The next step concerns our reaction to this phenomenon.

2. Every person observes this behavior, and each tries to articulate an explanation of each instance of behavior by assimilating it to the laws of agency he or she believes to obtain.

We often have ulterior motives in understanding other people, but in many cases this reaction is nearly automatic. If I see Jones pull a ripe apple out of the refrigerator and start eating it, I infer that he did so to sate his hunger. And I do this without any reflection on what I have to gain from this understanding. In part this is because offering an explanation is so easy (if it

33 4:437.
took real detective work to understand Jones’s snacking, I might have given up), but it is also because such explanation is an intellectual reflex. Indeed, a disposition to make sense of others is a condition on our occupying a common social sphere.34

The laws we turn to in seeking this understanding are not etched in stone. Sometimes we will revise our picture of the laws of agency if we can better explain the phenomena. That is the third step.

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

This step is an instance of the truism that we should revise our theories in the face of recalcitrant data. The details of how and when we revise our theory of agency are messy, but so they are in all cases of theory change. Suppose an interpreter has a theory of agency according to which performing a certain action from a certain motive is 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 the putative action genuinely uninterpretable and resort to non-interpretative explanation. But what if our interpreter keeps seeing this kind of action? What if he sees others acting as if they understood this kind of action perfectly well? Like any other scientist, at a certain point, our interpreter must stop explaining away this data and see it as genuinely recalcitrant. What had 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 constitutes the phenomena that those laws are meant to explain.

To this point, each step in this process has been a special case of a general principle about explanation. We have some phenomena, we try to explain them with the theory we have in hand, and we modify our theory if we think doing so will get us a better explanation. The next step is what distinguishes this process from other kinds of explanation.

34 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 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.
4. Every person’s theory of agency regulates her behavior through comparisons of potential courses of action to what that person takes the laws of agency to be.

The basic thought behind this step can be brought out by our analogy with chess. I am a novice chess player, and momentarily think that moving my bishop to attack my opponent’s queen is a winning strategy, but I then check this move against the rules of chess and discover that it is illegal. Of course, the ways that the “rules” of agency regulate action in general are not nearly so straightforward. We do not often hear the refrain, “doing this will make me less of an agent, so I shall abstain.” Instead we find subtler forms of regulation. One relatively explicit example is that of people being turned away from a plan by the fear that, despite its attractions, pursuing it would mean that they had gone mad or lost control of themselves. We find this in cases of self-signaling: actions we perform because we think their performance says something about us.35 Lenore will not have me, and yet I am ready to weather all the pains I might suffer in continuing to pursue her because I think that her love would be such a sublime thing. The only thing that keeps my sonnets off Lenore’s doorstep is the startling revelation that if I saw someone else doing as I plan to do, I would think he was out of his wits, that he had become so overwrought that he was no longer a reasonable judge of his interests. Similarly, suppose I am enchanted by the barkeep’s suggestion of another belt of whiskey and am angered by the softest whisper about last call. The only thing that stops my drinking is the reflection that these are the thoughts of an alcoholic, someone whose evaluative and motivational powers have been compromised by addiction. In these cases I am motivated by a fear of “losing myself”—by the thought that if I go down a certain path I will become something less than an agent. A theory of agency can also regulate behavior in more covert ways. Most importantly, it presents us with a menu of possible ways of being. It dictates the types of action available, the relationships that one can occupy, the kinds of character one can have, the palette of feelings one can experience. Ian Hacking offers a useful example of how a change in the theory of agency—the introduction of multiple personality theory—influenced the sorts of actions available to agents. “Instead of mood swings there was something much more specific that a person could be doing, namely, switching to a persecutor alter; a persecutor alter could be taking executive control. I do not mean that doubles could not switch, or that an alter could not come out before words were available…. We are content to apply such descriptions

35 There is very little philosophical work on self-signaling, but see Bodner and Prelec (2003).

retroactively to times when the descriptions were not available. . . . But it is not clear that the vivacious alter could intentionally come out, could choose to come out [before the development of multiple personality theory].”36 This example shows that what we do is influenced by the range of actions we think an agent can perform, and this is something we find in our theory of agency. The same is true of affective attitudes and character. We are not born with concepts like resentment, pique, wanderlust, and ennui. We have to learn them by observing others, and so my ability to resent Jones depends on my theory of agency recognizing such a thing as resentment. Similarly, I can try to turn myself into a Byronic hero because it is a character type I recognize in the world. (Not a common one, I admit.) Beyond making certain actions, affects, and character types possible, our theory of agency also clusters these things together in ways that affect our deliberations. I may love the way breeches accentuate my calf muscles, but according to my theory of agency, breech-wearing is something dandies do, and that’s the last thing I want to be. What we see in all these examples is that this crucial step in the Mutual Interpretation Process is accomplished not by a single channel of conscious regulation, but by a thousand rivulets of inconspicuous influence.

In the next step, the process begins again.

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

Rendering the Mutual Interpretation Process as a sequence of discrete steps is an idealization. In practice, the whole thing is a blur. Everyone is engaged in every step at every moment, theory revision is largely automatic and unconscious, deliberation does not involve any active comparisons of one’s actions to a theory of agency, and each person is unaware of the behavior of most other individuals.

But all this idealization serves a purpose. It reveals crucial interdependencies. Green’s actions depend on his understanding of agency, which in turn indirectly depends on Red’s actions (because Red’s actions are part of the explanandum of Green’s theory of agency) which in turn depend on Red’s understanding of agency (because that theory regulates his behavior), which depends on the actions of Green, Brown, Blue, and many, many others. One way to characterize this network of interdependencies is to say, à la Quine, that Green’s actions, Green’s theory of agency, Red’s actions, Red’s theory of agency, Brown’s actions, Brown’s theory of agency, and so on . . . all form a holistic web. Anything affecting one part of this web will be

communicated through these connections until it has affected all the rest. Another characterization comes from sociology. According to Anthony Giddens’s theory of **structuration**, actions are only possible in the midst of social structures that give them meaning. But Giddens does not think these structures are immutable. They are the dynamic creations of agents, the manifestations of our actions. Social structure and agency are mutually dependent: what we can do depends on the social structure, and the social structure depends on what we do. This relationship is an example of what Hacking calls a looping effect: “Think what the category of genius did to those Romantics who saw themselves as geniuses, and what their behavior did in turn to the category of genius itself. Think about the transformations effected by the notions of fat, overweight, anorexic. If someone talks about the social construction of genius or anorexia, they are likely talking about the idea, the individuals falling under the idea, the interaction between the idea and the people, and the manifold of social practices and institutions that these interactions involve.”

This mutual dependence is what distinguishes agency from other phenomena. 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 hopes of getting a different theory because that would mean hiding something from ourselves that is there anyway. But agency is different. We produce the phenomena that are to be explained with the laws of agency. Our actions are the essential constituents of the target phenomena, the thing a theory of agency is about. This arrangement scrambles the usual distinction between directions of fit. It also means that getting it right about agency requires something altogether different from getting it right about physics. What the Mutual Interpretation Process aims for is not the accurate reflection of an independent reality, but an **equilibrium** between its two sides, between our actions and our explanations of agency. (We could say between its “world-to-mind” side and its “mind-to-world” side.) Such an equilibrium will be a state in which 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, including herself, *qua* actor. At this state we will see two crucial kinds of alignment: amongst different individuals’ respective theories of agency, and between each of these theories of agency and the behavior present in the world.

---

This equilibrium is the aim of the Mutual Interpretation Process. It is a fixed point at which neither of the revision mechanisms included in the Mutual Interpretation Process—the revision of theory in light of evidence and of behavior in light of theory—forces movement in the overall system. Not all instances of the Mutual Interpretation Process will inexorably tend toward this equilibrium. But a society must approximate it for there to be any agency at all. Total equilibrium may not be necessary for the existence of agency. But too much discord amongst theories of agency or too much noise between theory and action destabilizes would-be agency to the point where it is impossible for it to have a definite constitution. When there is such an equilibrium, or something stable enough to approximate it, agency is successfully constituted.

With the workings of the Mutual Interpretation Process laid out thus, let me summarize where we stand and where we are headed. 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 introduce the dual processes of acting and interpreting into a social medium, one in which everyone else is doing these same two things, the result is the Mutual Interpretation Process. Now I shall argue that the Mutual Interpretation Process simply is the Realm of Ends procedure by a different name. We can discern the equivalence in three steps.

First, acting on a principle is, in effect, an attempt to legislate that principle as a law for the entire community because doing so is ipso facto to introduce the corresponding behavioral phenomenon into the world. This action exerts an influence over others because their respective accounts of agency must grapple with this new phenomenon. This influence is small and indirect, but, when accumulated over the many actions all of us 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 this action into the world of behavior as a sensible thing for a man to do, given his background beliefs and desires. This amounts to proposing a practical principle in Kant’s legislature to the effect that all those who have these background beliefs and desires ought to, other things being equal, ride mechanical bulls.

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 (though this is not entirely at their discretion), and they can hammer their verdict home by acting on principles incompatible with Tex’s. They
can act on practical principles that undercut the whole idea that bull-riding is a sensible way to woo a lady. Doing so introduces behavior into the world that places demands on our theory of the laws of agency incompatible with the demands Tex’s behavior imposes. This is tantamount to introducing a resolution to “table” Tex’s bill in the legislature. The more of this countervailing behavior there is, the greater the pressure on our theory of agency to understand Tex as an anomaly, as a crazy person whose bull-riding is a kind of pathology. It is therefore through the actions that we perform and the ensuing empirical pressure they exert on theories of agency that practical principles are negotiated.

Finally, the state of equilibrium that the Mutual Interpretation Process seeks is the very same state as the total agreement of practical principles that defines the Realm of Ends. The Realm of Ends is the equilibrium state of Kant’s legislature, the state in which there is agreement amongst legislators on practical principles and between those legislators and their actions. The equivalence follows immediately from our previous two claims. If two procedures are identical, then so are their equilibrium points. So the Realm of Ends is the Mutual Interpretation Process’s equilibrium point.

The requirement to adhere to the laws of agency, when combined with the fact that agency is jointly constituted, manifests itself as a requirement to engage in the Mutual Interpretation Process and thereby, to work toward that process’s equilibrium point. As we just saw, the Mutual Interpretation Process is identical to the legislative procedure of the Formula of the Realm of Ends, as are their respective equilibria. Therefore, the basic constitutivist thought that the requirements of agency correspond to categorical norms gives us:

*Realm of Ends Constitutivism.* We are categorically required to (a) participate in the contractualist procedure of the Formula of the Realm of Ends, and so to work toward the equilibrium point that is the Realm of Ends, and (b) adhere to the norms that emerge from any adequately stable approximation of this equilibrium.

To restate the whole argument: An individual is obligated to submit herself to this 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 Formula of the Realm of Ends.40

40 The idea that our *most basic* norms may turn on implicit negotiation goes back to Adam Smith. (Hobbes’s theory, of course, assumed more basic norms involving self-preservation.) In the *Theory of 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 says, these two forces regulate our sentiments and, indirectly, our actions. The

10. PROBLEMATIC CONSTITUTIONS

My view could be called “social constitutivism” because it sees agency as importantly (though not wholly) socially constructed, and it maintains that categorical norms arise from the social structures that constitute agency. Social constitutivism offers a striking advantage when we come to the question of whether the norms of agency will have a moral hue, whether, that is, we will owe anything to each other. Because agency is social, living up to its demands requires some form of cooperation. What kind of cooperation will vary from society to society, but every case will involve deference to others. This is far from a guarantee that common-sense morality will be categorical. (In fact, I doubt this very much.) But it is the leading edge of an argument for the categoricity of moralish requirements.

The virtues of this arrangement become clearer when compared to Korsgaard’s constitutivism. She maintains that categorical normativity emerges from the demands of self-constitution. This is a highly individualistic form of constitutivism, and for this reason it is difficult for her to show that we are required to take certain attitudes toward other people. My successful self-constitution seems compatible with many different attitudes toward other people, some kind, some very nasty.41 G. A. Cohen makes the point with a vivid example: an “idealized” Mafioso who is in perfect psychological equilibrium, reflectively endorses his vicious Mafioso ways, and accepts the Mafia code as his true practical identity. This fellow seems to have done quite well by the requirements of self-constitution, and yet we would hope that there is some censure we can hang on him.42

Social constitutivism offers a way to do this. The Mafioso’s violent outbursts at minor provocations are unintelligible to us; they are “senseless violence.”43 This reaction indicates that the Mafioso is that much less of an

41 Korsgaard recognizes this problem, and has tried to solve it in a few ways. In her (2009) she talks about fusing wills, an idea found in Kant’s writings on marriage. As far as I can tell, she argues that we can never be quite sure with whom we will have to unify our will, and so we have a general duty to respect all others arising from our duty to respect our own will. It is hard to see how wide this suggestion extends the circle of moral concern, but I doubt it is very far.


43 I am skating over the important issue of who this “us” is. Presumably if it is just the Mafioso’s compatriots, they will find his behavior intelligible under a Mafioso
agent, however self-satisfied he may be. (Of course, it does not vitiate his agency entirely: we can understand him as trying to live up to a code of conduct that we find partially intelligible, and this may make him more intelligible than someone who is completely incoherent.) He has therefore violated a categorical norm. On my view, our judgments about the Mafioso’s iniquity affect him because his behavior is what it is only as part of a matrix of public interpretation.44

For all that, my view is afflicted with its own socialized version of the Mafioso problem. The Mafioso case bids us imagine a person who satisfies all of Korsgaard’s criteria about coherence and reflective endorsement, but nonetheless runs afoul of some very obvious moral requirement. The corresponding burr under my saddle is the possibility of a society that approximates the equilibrium of the Mutual Interpretation Process, and so satisfies the conditions I presented for a successful constitution of agency, but through this constitution supports norms that are, intuitively speaking, wicked. Social constitutivism threatens to make normativity too beholden to actual social structures—too conservative—just as Korsgaard’s threatens to make it too beholden to actual psychology. We may be able to criticize individuals like our Mafioso for stepping outside of agency’s social contract, but we have no resources to criticize stable constitutions of agency, no matter how morally unworthy they may seem.

My answer to this objection is that many constitutions of agency that appear stable are not, or at least not perfectly so. These instabilities represent ways in which actors could come en masse to withhold their assent from the legislation of practical principles. They are ways in which a constitution of agency could fall apart. These instabilities are sometimes hard to see: they can be obscured by an entrenched ideology (that one group of people is naturally subordinated to another), suppressed constitution of agency. So I am assuming that it reaches further than that. But whom should it include? All and only those who live in the same city as the Mafioso? All and only his countrymen? All and only his contemporaries? (Does our inability to interpret Caligula matter to his agency?) These are questions I cannot answer here.

44 The Mafioso’s case is an interesting one for my account because the intelligibility of his actions is no doubt affected by the image of the mobster in the popular imagination, which in turn is affected by the ubiquitous portrayal of organized crime in popular media. The pulpier portrayals may romanticize the Mafioso mindset and glamorize his crimes. But more finely honed examples of the genre can reveal cracks and fissures lying beneath the façade of the Mafioso mindset. The stoical and efficient assassin played by Alain Delon in Jean-Pierre Melville’s Le Samourai is precisely the kind of character that Cohen’s example brings to mind, but in Melville’s film he is portrayed as hollow and inhuman. The genius of the film is that its central mystery is not how some caper is to be pulled off, but why its protagonist does what he does.
by homeostatic mechanisms (legal strictures and economic incentives that encourage conformity to a given constitution), or simply hidden amid the din.45

Consider two simple examples of social constitutions that appear stable but have obvious moral shortcomings: the Jim Crow South and the factory culture of the Gilded Age. The very features of these societies that made wicked things permissible, even obligatory, were also sources of instability. The abuse of workers was so thoroughly ingrained in the psychology of Gilded Age managers like Henry Clay Frick and their underlings that scarcely any tactic for increasing profits, no matter how brutal, was beyond comprehension. Shooting strikers, bribing politicians, and chaining workers to their desks were all intelligible business strategies. They were part of the constitution of agency. Because there was no check on the heinousness of worker treatment, abuse spiraled out of control and ultimately precipitated sensational events like the Triangle Shirtwaist Factory fire. While smaller contretemps may have been overlooked by politicians, the wider public, and other workers, the factory system could not survive such a spectacle. Thus it was the unmitigated cruelty inherent in factory system’s constitution of agency that sowed the seeds of that constitution’s collapse. We see the same thing in the Jim Crow South. The racial oppression of that time and place was rooted in what W. J. Cash famously called a “proto-Dorian bond.” The proto-Dorian bond was a crucial part of the constitution of agency in the South. It bound whites together in the idea that the most fundamental division of society was racial, and that this division was rooted in “natural” differences between the races.46 This piece of ideology stabilized racial oppression in the face of many challenges. Segregation was justified by the differences between the races, and these

45 The concept of stability does not quite capture what I have in mind in this reply. A more complete way to put this point would employ the Kantian idea of systematicity. Systematicity is a regulative ideal imposed by reason. It is something we can never fully attain in our projects—both theoretical and practical—but it nonetheless regulates our projects by acting as a model we work toward. For theoretical reason, this ideal is not too different from the ideal of “beauty” sometimes invoked as a reason to prefer one explanation to another. (Kant offers a definition of systematicity beginning around A645/B673; see Paul Guyer (2005, pp. 245 ff.) for commentary.) The particular form this ideal takes for practical reason is the Realm of Ends. The Realm of Ends is a perfectly systematic practical organization of wills: orderly, symmetrical, unified. It has all the structural features that would merit the attribution of beauty in an explanation. Stability, I maintain, is an important facet of systematicity (though perhaps a derivative one), but systematicity demands more than stability. Here I talk about stability instead of systematicity because an adequate explication and defense of Kant’s ideas about systematicity would be too large an undertaking. Stability is a more familiar concept, and it is close enough for me to make my point.
46 Cash (1941).
differences were rooted in the nature of the races. So calls for the destruction of segregation were rooted in mistakes about the workings of the world. But the centrality of the proto-Dorian bond to Jim Crow agency was also a source of instability. Claims about the “naturalness” of segregation were undermined by scholarship showing the contingent and relatively recent origins of the Jim Crow system. Jim Crow was not an antediluvian way of life. It was the kind of system that Giddens identified: a structure that imposed genuine constraints on people, but was also created by the actions of those people. The emergence of this scholarship changed the way blacks and whites thought about themselves, and in doing so arrested some of the power of the proto-Dorian bond. With this change in view, the overt parts of the Jim Crow apparatus—de jure segregation and organized violence—stopped being seen as necessary mechanisms for the maintenance of social order and were revealed rather as instruments of exploitation and oppression. This was an essential step in the eventual collapse of the Jim Crow system. I have presented these examples as just-so histories, but we see the same pattern when we turn to toy examples. We may worry about a world of knaves legislating a knavish constitution of agency through their knavery. Would my view require these knaves to persist in their knavery? The closest thing to this world I can conjure up is the setting of Dashiell Hammett’s novel Red Harvest. The Montana town of Personville is populated by gangsters, thugs, murderers, drunks, gamblers, and hangers-on who feud constantly, steal without compunction, and murder for pocket change. These anti-social behaviors are so entrenched that they constitute a form of agency according to which it makes sense to kill a man over a dollar; indeed it would be unintelligible to turn the other cheek. Personville is the world nearest a Realm of Knavery I can imagine. But the whole point of Hammett’s novel

47 Most prominent in this respect is C. Vann Woodward’s 1955 book The Strange Career of Jim Crow, which had an enormous influence over the leaders of the Civil Rights movement.

48 The example of the Jim Crow South is an instructive one. Revolutions of agency are usually precipitated by two seminal realizations. First, agents come to realize that a given constitution of agency is not “natural” in the sense of the proto-Dorian bond, but artificial in Giddens’s sense; it is dependent on the actions of agents. This means that this constitution of agency is, in a way, optional. One cannot simply flout it—it does have normative force—but it is capable of reform by collective action. Second, these revolutionaries come to realize that the consent of many agents to this constitution of agency (to the legislation of practical principles) is premised on a mistake: about themselves, about their interests, or about the order of the world. The kind of mistake I have in mind is similar to what Marxists call “false consciousness.” Bringing to light these two types of fact is a central goal of a critical theory. See Horkheimer (1937).
is that such a constitution is inherently unstable. The residents of Personville do agree on violence and thievery; that is how they manage to constitute a knavish form of agency. But they are fundamentally and violently at odds with each other, a fact that leaves this form of agency vulnerable. So when a clever and adroit man drifts into town, someone like Hammett’s protagonist, he can exploit this system. He can play one gang of knaves against the other and draw them into a struggle that leaves many of them dead and shatters their constitution of agency. At the end of the novel Hammett even offers a remedy for this instability, one straight out of Hobbes: the old industrial baron is restored to his place as Sovereign.

In each of these examples we see that the very things we would call wicked, immoral, or otherwise problematic are also sources of instability. This will not always be the case. Realm of Ends Constitutivism is an emphatically revisionary theory of normativity, and it will not agree with all of our moral intuitions. Nonetheless, the constitutions of agency we are particularly worried about supporting iniquitous norms because they involve oppression, exploitation, or wanton violence are unstable precisely because they involve these things.

I do not mean to deny that these unstable constitutions produce categorical norms. I said that any social organization that adequately approximated the equilibrium of the Mutual Interpretation Process (that is, the Realm of Ends) had normative force, and these constitutions qualify. So I accept the consequence of my view that some find objectionable: the denizens of these societies really ought to behave in conformity with their particular constitutions of agency because that is what it takes for them to act. What the instability of the Jim Crow South, Gilded Age factory culture, and Personville shows is that this need not be the final word.

According to Realm of Ends Constitutivism we are required to participate in a contractualist procedure. Out of this procedure come two different kinds of norms. On the one hand there are norms that emerge from a “stable enough” contract of practical principles, that is, norms associated with a constitution of agency that is stable enough to persist. On the other hand, there is a more global norm requiring one to continue participating in this contractualist procedure, even when one has reached a “stable enough” constitution of agency. The Jim Crow South and Personville may possess “stable enough” constitutions of agency, but they are imperfect. And so long as there are more perfect constitutions available, the Mutual Interpretation Process carries on.

The thing toward which these people must work is the Realm of Ends. The Realm of Ends is the ideal conclusion of the Mutual Interpretation
Process. It is the state that most perfectly lives up to the demands inherent in that process. The Realm of Ends is ideal in two different senses. First, it guides the Mutual Interpretation Process by giving us a model to strive for. Second, it is unattainable. There is a tendency to understand the Realm of Ends as the “liberal’s idea of a good society” (to quote R. M. Hare). But this is a mistake. It is something more distant. 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.”

Our two kinds of norm will very often be in conflict with each other. The norm urging us closer to the Realm of Ends will have us break down extant social structures so they may be replaced with something better, while the norms arising from particular “stable enough” constitutions of agency will require actions whose effect will be the perpetuation of those structures. I cannot begin to sketch an attempt of how these norms interact, but I will mention something interesting that this dualism entails: a theory of normative change. Normative revolutions are fomented by behavior that is at the margins of intelligibility. Behavior that is too far out of step with the reigning constitution of agency will be dismissed as lunacy, but someone whose behavior mostly conforms to these standards will be an agent (if an imperfect one). Not any violation will produce a revolution in our constitution of agency. For this to happen, an action must provoke critical reflection on the nature of agency. In particular, it must lead observers to three essential realizations: that the constitution of agency is mutually dependent on the actions of human beings, and to that extent a particular constitution of agency is optional; that many agents’ consent to the legislation of practical principles that produces this constitution is premised on a mistake (for instance, the false belief that the segregation of blacks and whites is something rooted in their different natures); and that a more perfect constitution of agency, one nearer the Realm of Ends, is possible. The behavior of moral revolutionaries can communicate these ideas by exemplifying an alternative conception of agency.

49 See my previous footnote 45 about the ideal of systematicity. The Realm of Ends is the regulative ideal of systematicity for practical reason, and it is also the regulative ideal of systematicity for theoretical reason when it attempts to make sense of the emergent property of agency.


51 I cannot offer an account of exemplification here, other than to say I have Nelson Goodman’s (1976) notion in mind. For a really fascinating account of philosophers thinking of themselves in these terms, see Ineke Sluiter’s (2005) discussion of the Cynics’ philosophical method.
of itself that it is a proper action, and when it does this in conscious disagreement with the reigning constitution of agency, this assertion amounts to a suggestion that that constitution is optional. We see this dynamic most clearly in performance artists, perhaps even in Lady Gaga. By doing outlandish things, performance artists can focus our attention on the limitations of agency, and, in the appropriate context, thereby suggest that these limitations are artificial and arbitrary. More powerful examples come from the annals of civil disobedience. Rosa Parks’s refusal to give up her seat on the bus was not just a refusal. Her action exemplified an alternative conception of agency, one whose fundamental axiom was not the separateness of the races. In the political climate in which it was performed, Parks’s action communicated two things: that there are other ways to be an agent—meaning that Jim Crow agency was optional and required justification—and that her competing constitution of agency better lived up to our ideals of justice.

11. CONCLUSION

I began with the challenge of formulating a theory of normativity that avoided the pitfalls of both realism and Kant’s constitutivism. In answer to this challenge I proposed that we needed a constitutivist theory built on a conception of agency that was neither independent of the world (as Kant’s was) nor reducible to physics. We needed a conception of agency as an emergent property. We next saw that agency is a very special kind of emergent property. It is one whose nature is partially fixed by us through the interaction of interpretation and action that I called the Mutual Interpretation Process. This fact had the consequence that normativity is, in a sense, socially constructed, and that abiding by the demands of agency means submitting oneself to the contractualist procedure found in Kant’s Formula of the Realm of Ends. 52

52 I am grateful to Sally Haslanger, Richard Holton, Sophie Horowitz, Rae Langton, Victor Kumar, Paulina Sliwa, and two anonymous reviewers for comments on earlier versions of this essay. Thanks also to audiences at the MIT Ethics Group, the Wisconsin Metaethics Workshop, and Temple University for questions and helpful suggestions that improved the essay. I am especially grateful to Kate Elgin, Julia Markovits, and Alice Phillips Walden for exceptionally generous comments.
REFERENCES


Author Queries

[AQ1] Running head not fitted. So we have shorten the running head. Please check