**Ergo** AN OPEN ACCESS JOURNAL OF PHILOSOPHY

# Morality, Agency, and Other People

KENNETH WALDEN

Constitutivists believe that we can derive universally and unconditionally authoritative norms from the conditions of agency. Thus if *c* is a condition of agency, then you ought to live in conformity with *c* no matter what your particular ends, projects, or station. Much has been said about the validity of the inference, but that's not my topic here. I want to assume it is valid and talk about what I take to be the highest ambition of constitutivism: the prospect of grounding moral requirements in the conditions of agency. If this can be done, then we can show that everyone is bound by the demands of morality, and we can do so without the customary entanglements—queer normative entities, an implausibly powerful moral sense, or divine lawgivers.

Kant had this ambition (on one reading of his moral metaphysics, anyway). For him the moral law's universality meant that it had to be a law of freedom, a law that characterized the activity of autonomous wills. It was also the aspiration, in more complicated ways, of post-Kantians like Fichte, Hegel, and Bradley. And it is a project pursued by some contemporary philosophers. But there is something surprising about this final group's efforts. They begin with a conception of agency that appears highly individualistic, a conception whose conditions don't explicitly mention other people. This is surprising because presumably the goal of deriving the universal authority of moral requirements from a constitutivist argument will involve demonstrating that *other people* play some distinctive role in my agency—a role that requires me to honor, respect, or care for them. So if other people are not a party to my agency, it is hard to see how we are supposed to establish this sort of conclusion.

Looking back, we find that claims about the "sociality" of agency enter as key premises in the arguments of historical constitutivists. In Kant we find a tight connection between personal autonomy and the ideal of interpersonal unity enshrined in the Realm of Ends. We are only capable of acting on a law we create for ourselves, Kant says, if that law is freely adopted by other legisla-

**Contact:** Kenneth Walden <kenneth.e.walden@dartmouth.edu>

http://dx.doi.org/10.3998/ergo.12405314.0005.003 

tors. This gives those legislators standing with respect to my practical reasoning.[1] The conditions of "self-positing" play a similar role for Fichte. One such condition is positing myself as an *individual*. To do this Fichte thinks I must recognize other persons as free selves, which involves acknowledging their "summons" and eventually coming to limit my own freedom out of respect for theirs (2000: 31ff.). In the *Phenomenology of Spirit* Hegel says that "self-consciousness can only achieve satisfaction in another self-consciousness" (1807/1977: §175), which suggests that a form of mutual recognition is a condition on all action that can be plausibly called self-governed. Finally, Bradley (1879) argues that our true and authentic self is a social self, constituted by our "station" in society, and thus in "realizing" our self through our actions we must live according to that station.

Whatever the plausibility of these claims, we can at least appreciate why someone interested in grounding the demands of morality in the conditions of agency would be pulled in their direction. It's surprising then that we don't find similar theses about the essential sociality of agency defended by latter-day constitutivists. Or so I shall argue. In this paper, I flesh out this concern by looking at two leading constitutivist moral apologetics, those of Christine M. Korsgaard and J. David Velleman. Korsgaard believes we can extract a duty to respect other persons as ends in themselves from the demands of agency. Velleman's project is more circumspect. His claim is not that particular moral requirements can be derived from the constitutive aims of agency, but that those aims "push" us in the direction of "our moral way of life". In what follows I will argue that Korsgaard's argument fails and that Velleman's, even with these provisos, doesn't secure the right kind of vindication. Both of these weaknesses, I will suggest, are symptoms of setting off from an individualistic conception of agency.

My second task is the advancement of a thesis about agency that I think will serve as a better foundation for a constitutivist validation of morality. My suggestion is that a person's agency depends on her interpretability by others in a way that obliges her to recognize a distinctive authority in other persons. No lengthy canon of moral duties falls directly out of this conclusion, but I argue that we can nevertheless derive a synoptic duty of respect for persons from it.[2]

1. See Reath (2006).

2. In other work I have made a related argument. I argue that agency has social conditions because it is an "interactive kind". These conditions commit agents not to any particular kinds of behavior, but to a process of negotiation about the nature of the kind. This negotiation ends up being identical to the legislation that defines the Realm of Ends. I see these arguments as related in roughly the way that Kant's arguments for the Formulae of Humanity and the Realm of Ends are. See Walden (2012).

## 1. Korsgaard's Heroic Argument

According to Korsgaard, when we act we are trying to constitute ourselves as unified agents. This means that self-constitution is a constitutive aim of all action. By the logic of constitutivism, it follows that the demands of self-constitution have a special normative authority. Korsgaard's strategy is to show that particular principles are normative because it is only by following them that we can unify ourselves as agents. This suggests that for her the site of self-constitution is local; it is one's own self. Whether I am successfully constituted is a matter of my parts fitting together into a stable, integrated, and harmonious union. But this does not depend on my relationship to anyone else, at least not obviously. I may choose to make my self-constitution dependent on Felix by falling in love with him. And if Felix turns out to hate my religion, then this is may be a real threat to my constitution. But on the face of it, there is nothing about self-constitution *per se* that requires entangling oneself with others.

Korsgaard's derivation of particular norms bears out this assessment. She defends the normativity of two principles on the basis of their connection to our self-constitution as unified agents. We must follow the hypothetical imperative to be efficacious and the categorical imperative to be autonomous, but efficacy and autonomy are themselves required for successfully constituting oneself as an agent, and that makes these principles normative:

> The ideal of agency is the ideal of inserting yourself into the causal order, in such a way as to make a genuine difference in the world. Autonomy, in Kant's sense of not being determined by an alien cause, and efficacy, in the sense of making a difference in the world that is genuinely your own, are just the two faces of that ideal, one looking behind, and the other looking forward. That is why Kant's two imperatives together are the laws of agency. (Korsgaard 2009: 89–90)

The ideal Korsgaard describes here is highly individualistic. To be an agent in this sense is to succeed in "inserting yourself in the causal order in such a way as to make a genuine difference in the world" and that seems to be a project whose success or failure depends entirely on how things are with *you* and the part of the causal nexus you inhabit, and not necessarily on your relations with other people. For this reason, it seems doubtful that we will be able to locate other-regarding principles—including moral ones—in the requirements of agency as Korsgaard lays them out.

But, one might object, Korsgaard claims to have grounded the categorical imperative in the conditions of agency. And isn't that a moral principle? This label is a little misleading, as Korsgaard herself explains:

> In *The Sources of Normativity*, I distinguished what I called "the categorical imperative'' from what I called ''the moral law". The categorical imperative is the law of acting only on maxims that you can will to be universal laws. The moral law, as I characterized it there, is the law of acting only on maxims that all rational beings could act on together in a workable cooperative system. The arguments I've given above don't—or rather don't obviously—get us all the way to a commitment to the moral law in that more specific sense. To get from the categorical imperative to the moral law, two more things are necessary. First of all, we must establish that the domain over which the universal law ranges must be rational beings as such: that is to say, when you will your maxim as a universal law, you must will it as a law for every rational being. And second, we must establish that the reasons embodied in universal maxims must be understood as public, or shareable reasons: reasons that have normative force for all rational beings. These are issues to which we will return in Chapter 9. (Korsgaard 2009: 80)

Chapter 9 is pivotal for Korsgaard's project, then. It is the place where she aims to show that the ideal of agency she describes, which I have suggested looks highly individualistic, actually binds us to other people. Moreover, she hopes to show that these interpersonal bonds subject everyone to an indisputably moral principle: respect for the humanity of persons. The remainder of this section is devoted to reconstructing and critiquing this argument.[3]

Korsgaard begins her argument with an analysis of joint action. She says performing a joint action requires joint deliberation, and this, in turn, requires a fusion of agency.

> To perform a shared action, each of us has to adopt the other's reasons as her own, that is, as normative considerations with a bearing on her own case. . . . The aim of the shared deliberation, the deliberation about when to meet, is to find (or construct) a shared good, the object of our unified will, which we then pursue by a shared action. And it follows from the fact that the action is shared that if either of us fails to show up, we will

3. In earlier work Korsgaard has made different arguments with the aim of showing that what appear to be self-regarding normative commitments are in fact other-regarding. In the final chapter of *The Sources of Normativity* (1996) she gives an argument for the publicity of reasons modeled after Wittgenstein's argument for the publicity of language. One could, in principle, supplement Korsgaard's account of self-constitution with *that* argument instead of with the one she goes on to make in chapter 9 of *Self-Constitution* (2009). I don't deal with this possibility because the argument surveyed here is a more natural complement to the constitutivism developed earlier in the book and the Wittgensteinian argument has already attracted substantial critical attention, e.g., in Joshua Gert (2002: 303–324) and R. Jay Wallace (2009: 471–498).

> both have failed to do what we set out to do. Our autonomy and our efficacy stand or fall together. (Korsgaard 2009: 192)

It follows from this that the demands of self-constitution may span the gulf between agents. When we enter into this kind of deliberation, these demands become principles of *joint* constitution. And this, in effect, requires each joint deliberator to treat her comrade with respect. Just as I must treat myself as an end in myself because I am the seat of practical reasoning, I must treat my partner in the same way because she is part of the practical reasoning we share.

What drives this argument is the assumption that our two parties have embarked on a strong form of interaction: joint deliberation *en route* to joint action. Someone could resist the putative obligation to respect another person simply by declining to interact with them in this way. This need not entail any kind of radical sequestration. It is not just the hermit who never interacts with people in this way that is untouched by the conclusion. Someone who has a policy of breaking off this kind of joint deliberation the moment it becomes advantageous to use or mistreat an erstwhile comrade would be doing nothing wrong. Thus saying that within the fold of joint deliberation we must treat others as ends in themselves is not to say very much. For this kind of argument to succeed, we need reasons to enter into and persist in these strongly collaborative relationships. Korsgaard acknowledges this possibility and responds by suggesting a kind of interaction that we cannot eschew.

> Let's suppose that you can just decide to treat someone's reasons as reasons, with normative implications for you, but that you need not do that unless you choose to. Would that show that morality is optional, depending as it does on whether you have any private reasons that favor personal interaction? It is not that simple, for there is one person with whom even the most determined private reasoner must interact in the way that Kant's theory requires. And that is himself. (Korsgaard 2009: 202)

I must interact with myself, Korsgaard suggests, in order to unify myself as an agent. In particular, I must interact with my later self to achieve diachronic unification, and so, per the previous thought, I cannot simply opt out of the requirement to respect other temporal fragments of myself.

This closes the loophole in the previous argument, but only by an inch. The fact that I am condemned to interact with my future self pulls that time slice into the ambit of respect, but no one else. Korsgaard raises this point herself and offers an intriguing reply:

> But couldn't [an agent] still will [a principle] as a public law only for him-

> self, binding together only the parts—parts of the soul, or time slices, or whatever they might be—that are parts of himself? Couldn't he, that is, decide to respect only his own humanity? This is an ill-formed question. What is your own, in the individual sense of your own, is not your humanity but what you make of it, your practical identity, and the existence of that *depends* on your respect for humanity in general. And besides—or maybe this is the same point—to respect your own humanity is to respect your own reasons, and we have already seen . . . that the category of ''my own reasons'' cannot be fully identified in advance of choice. (Korsgaard 2009: 204–205)

Korsgaard's thought seems to be that a policy of interacting with (and so respecting) only ourselves is "ill-formed" because the referent of "myself" is not specified in advance of action, and so in advance of action this policy lacks the content it needs to guide our deliberations. Thus a policy of limiting my interactions to only myself would be defective in roughly the same way that the intention of going two feet to the left of wherever I end up going is defective. They are both viciously circular insofar as their content depends on the performance of an action that they are supposed to produce.

Korsgaard doesn't say this, but I think she means—and needs—something stronger here. It's not just that an intention to interact only with *myself* is ill-formed, but that any intention restricting my interactions to *a particular set of selves* is ill-formed because all those selves are similarly constituted by action, and so not delimited in advance of acting. Without this stronger claim, it seems possible for me to restrict my interactions to myself and a small coterie of collaborators.

This is Korsgaard's last word on the subject, and a moment later she reflects on her argument by suggesting that it establishes that "respect for humanity is a necessary condition of effective action." So I think we can take her argument to be complete. I understand it thus:

1. To interact with another person I must respect her humanity.
2. I must interact with myself.
3. Any proposed restriction of my interactions to a particular set of selves (viz., my present self, my temporally self, my temporally extended self plus Felix's temporally extended self . . .) is ill-formed because these selves are constituted by action.
4. Therefore, I must respect the humanity of everyone.

This is an ingenious argument, but I don't think it is sound. A few problems deserve comment.

One might have doubts about the first premise. Why must I respect someone's humanity in order to interact with her? Can't I take a decidedly strategic stance toward her and still make a request, ask a question, or have a conversation? Korsgaard can respond to this objection by pointing out that she has a particularly strong conception of "interaction" here. As I understand her, interacting with another person means engaging in a joint activity. In her example, she imagines deliberating with a student as part of the performance of the action *meeting together*. That interaction in *this* sense presupposes respect seems more plausible. On this point, Margaret Gilbert's account of joint activity makes a useful supplement to Korsgaard's argument. Gilbert (2014) argues that joint activities are predicated on an (often tacit) joint commitment on the part of the activity's participants, and that this commitment grounds a "package of rights" that each party to the activity has against the other. One can disagree with this picture, of course, but if we accept it, then Korsgaard's first premise looks more plausible.

That said, such a gloss on the first premise puts considerably more pressure on the remainder of the argument, since it must now show that an agent must interact with everyone in a rather stronger sense than we might've supposed. And this is where we find more serious trouble. First, more needs to be said about premise (3). What is so problematic about using some point of reference that it is constituted by my actions as a guide to those actions? There *is* a kind of circularity here, but it's not obviously vicious. Suppose that we are homesteaders on the prairie. There aren't well-defined lots and property lines, but we all find our own spots and start building our houses. At one point my neighbor comes and suggests that we help each other with our houses, maybe even put up a duplex. By Korsgaard's lights, there would be something "ill-formed" in saying that I would prefer to work on just my house. I don't yet have a house—I've just begun—and the question at issue is exactly how my neighbor and I will constitute our houses. So my declaration seems problematic in precisely the way that Korsgaard says an intention to interact only with myself is. But in practice we don't seem to face this kind of problem. I have a practicable enough sense of what qualifies as my house and what I must do to keep it from "interacting" with my neighbor's house, even though the house is still very much in progress. The same seems to be true of selves. It may be that my life is a project of self-constitution, that at any point this project is still on-going, and that it's an open question whom I interact with in the strong way Korsgaard says involves unification of agency. But this does not mean that I don't have enough of a sense of where I end and Felix begins to declare that I will never interact with Felix in the way relevant to Korsgaard's argument. I can rule out unifying my agency with Felix's without supposing that the boundaries of myself are fully fixed.

Finally, this argument has the same logical form as the Sorites paradox. It

involves an inference from the illegitimacy of any particular boundary around a class to a universal generalization. In the Sorites argument, this is the inference from the falsity of any boundary condition on a vague predicate to the conclusion that either all or no items fall under the predicate, for example, from the falsity of all claims of the form, "a person with $n$ hairs is bald, but a person with $n$+1 hairs is not bald" to the conclusion that either no one or everyone is bald. In Korsgaard's argument it is the inference from the illegitimacy of any restriction to the range of our interactions to the conclusion that we must be open to interacting with anyone. We should find this reasoning invalid in both cases. The fact that all potential boundaries on whom we interact with are "ill-formed" does not mean we must interact with absolutely everyone. Our sphere of interaction, in Korsgaard's sense, may instead be vague in the way that the predicate "bald" is. There may be no bright line, but there *is* nonetheless an inside and outside. Indeed, we might think that it is wrong to insist on any definite restriction to my sphere of interaction since that would clash with my autonomy to make myself as I see fit. But this is not incompatible with declaring that Felix is definitively *beyond* that sphere, just as the person with a quarter-million hairs is definitively beyond the extension of "bald".

If these objections are successful, then Korsgaard cannot get her constitutivist case for respect to extend beyond a small circle around myself, and her argument does not show that respect for humanity *tout court* is a requirement of effective action. And that leaves us with our initial reservations about the character of her constitutivism. Namely, Korsgaard's theory is *prima facie* poorly suited to grounding other-regarding norms—and so poorly suited to be a foundation of morality—because it focuses on features of agency that seem achievable in isolation.

## 2. Velleman's Kinda Kantianism

David Velleman offers a simple standard for action and agency. "Action is behavior aimed at intelligibility," he says, "just as belief is acceptance aimed at truth." (2009: 133) And for this reason the "criterion of correctness" for action is "intelligibility because intelligibility is its constitutive aim" (2009: 134). What Velleman means is that in acting we are trying to make sense of ourselves, to understand what we are doing in light of our motives and beliefs. We are like an improvisational actor who is trying to act in ways that make sense of the character he is enacting through his performance. And this fact gives us a normative correctness criterion: an action is successful insofar as it makes us intelligible to ourselves and unsuccessful insofar as it doesn't.

Naturally one can disagree with this claim about the constitutive aim of ac-

tion, but I want to grant it. Instead, I want to make the same observation I made about Korsgaard's constitutive claim. On the face of it, self-intelligibility is something that I can achieve alone, without the help of anyone else. And so by initial appearances Velleman's constitutivism is not going to be any more successful in grounding other-regarding norms than Korsgaard's.

I don't think Velleman would necessarily disagree with this because his ambitions are a notch lower than Korsgaard's. He doesn't think we can derive anything like the moral law from the constitutive aim of action. Instead, he pursues a "Kinda Kantian" strategy according to which this aim "pushes" us in the direction of "our moral way of life" (Velleman 2009: 149). It does this, he says, because the aim of self-understanding is well-served by forms of interaction that have features we might call proto-moral. For one, this aim "favors developing intrapersonally coherent and interpersonally shared values." And for agents who are interacting it "requires them to join in an improvisational collaboration, which is facilitated by adherence to socially shared scenarios." Moreover, these collaborations are "generally facilitated by mutual understandings and hindered by deception" and "recognizing one another as rational agents should inspire a complex interpersonal regard." And finally, "our participation in joint improvisation fosters the development of a discrete mental process that functions in various ways ordinarily associated with conscience." (Velleman 2009: 149–150) These arguments suggest, Velleman says,

> a rough configuration that our dealings together would acquire from practical reasoning in the very long run: shared values and scenarios, discouraging private exceptions, minimizing occasions for deception, shaped by acknowledged common interest in comprehensibility, consequently free of unnecessary distinctions among persons, and supported by a psychological process recognizable as the conscience. (2009: 151)[4]

Velleman is not out to derive morality from the demands of agency. In a certain sense he doesn't believe that there is such a thing as "morality". His claim is that our aim of self-understanding is better served when our interactions with other people take on this proto-moral form—when those interactions exhibit "universality, transparency, and mutuality" (2009: 161).

I will not challenge any part of this argument. Instead I want to suggest that what Velleman's Kinda Kantianism vindicates falls well short of what aspiring moral constitutivists should hope for. One crucial feature of Velleman's story is a two-tiered normative structure, something it shares with contractarian accounts

4. Velleman has since backed away from even this more modest project because he thinks there is too much variability in the agency amongst different societies. See Velleman (2015).

of the authority of morality. The first tier specifies norms inherent in a particular form of shared interaction. I ought not use my neighbor on the subway as a footrest because that's not how the interaction between subway riders goes. The second tier is an explanation of why these norms are genuinely binding on me, of why I have a reason to adhere to them. Like the contractarians, Velleman's explanation at this stage involves a self-regarding principle: adopting these norms is my best way of living up to the ideal of self-understanding in a social world. The key difference between Velleman and the contractarians, I think, is that he takes self-understanding to be the aim of action rather than, for example, self-preservation (as it is for Hobbes) or preference satisfaction (as it is for Gauthier).

This feature makes Velleman vulnerable to versions of the standard objections to contractarian accounts. First, the strength and direction of the "push" that the constitutive aim of action gives us seems contingent. If human psychology or culture had evolved differently, there may have been significant exceptions to the transparency, reciprocity, and conscientiousness that Velleman describes. It's hard to say what changes would prompt which exceptions, but it does seem likely that these features will have less modal robustness than we usually associate with morality. Second, on Velleman's account other persons will merit moral treatment only insofar as they participate in the scenarios that further our self-understanding. There are a few problems with this. It conflicts with conventional wisdom about what entitles a person to decent treatment. It also threatens us with parochialism, since it seems likely that relatively exclusive and insulated practices will better serve the aim of self-understanding than more inclusive and open ones. Third, because the seminal norm in this whole picture is self-directed, we cannot say that we have any obligations that are irreducibly *to* other individuals, but only that we have obligations to ourselves that are, as a matter of fact, best met by acting *as if* we had such obligations. Nor can we say that this story gives us a reason to respect other persons as ends in themselves—as Korsgaard's does—since proving useful in a scenario that enhances my self-understanding is not a reason to respect someone, but, at most, a reason to ape respect.[5]

My objection to Velleman's account is that it fails to vindicate crucial features of our common conception of morality. Of course, one can reply that this is the best we can do, that we are mistaken in demanding a metaethics that entails strong claims like the inherent dignity of all persons or an unconditional duty of respect. All we can do is show that moral practices, taken as a whole, will better satisfy the constitutive requirements of action than the alternatives. Whether this reply carries the day depends on whether these "strong claims" can be established by other means. In what follows I will try to show they can.

5. For similar objections lodged against Gauthier's contractarianism, see Nicholas Southwood (2010: 34–48).

## 3. Toward a Social Constitutivism

I have argued that Korsgaard and Velleman's constitutivist vindications of morality both fail. They fail in different ways: Korsgaard's heroic argument is not sound and Velleman's Kinda Kantian strategy ends up grounding an anemic version of morality. Ultimately, I think they fail for the same reason. Because they do not include other people as essential constituents of the conditions of agency, they cannot construct a constitutivist validation of the claim that we owe something to these other people as such. And this would seem to exclude the possibility of establishing the universal and unconditional authority of other-regarding norms.

We might think that this reflects an inevitable limitation on the constitutivist program. Whether I am an agent comes down to facts about me. I can start my car, shave my face, and make a cup of tea all by myself, and if I do enough of these things I am an agent. So it would seem that agency is something that supervenes on the would-be agent and not anyone else. Of course, I may choose to interact with others and once I do this, these others may come to affect my agency. I need Veronica's help to perform the action *taking Veronica to the prom*. But there is nothing about agency and action as such that engender any interpersonal entanglements. Robinson Crusoe is capable of full-blooded and unimpeded agency. If we accept this thought, then the lesson of the foregoing sections may be that even if constitutivists can tell us cogent stories about self-regarding norms like those of rationality or prudence, they cannot ground principles about what we owe each other.

I want to suggest otherwise. Acting is not, strictly speaking, something we can do alone, and this fact can be the leading edge of a better constitutivism about morality. To do this I will need to establish a thesis in the same ballpark as the ones from Kant, Fichte, Hegel, and Bradley that I mentioned before. The key to such a project is finding a way that other persons figure into our projects that is both inescapable and requires something recognizably moral.

The best way to proceed here, I think, is to argue that our agency is conditioned on our participation in an activity that is at once governed by proto-moral attitudes of mutual respect and yet so broad as to encompass every other activity we might engage in. If we can do this, then we can show that all our actions are answerable to the demands of respect. And that would establish the normative authority of a highly schematic sort of contractualism and vindicate a principle close to the Formula of Humanity.

I will try to do this by arguing that the activity of *mutual interpretation* has these features. Participation in this activity is a condition on agency because it is a condition on our possessing the intentional attitudes necessary for our behavior to qualify as actions. And our participation requires us to recognize a

distinctive authority in other persons. The remainder of the paper is devoted to this argument.

## 4. Agency, Intentionality, and Interpretability

The first premise of my argument is that for a creature to be an agent, it must have intentional attitudes. By "intentional" attitude I mean an attitude that is about something in the world, the way my belief that my tea cup is half full is about that cup and my desire to drink tea is about a kind of dirty water. Agency requires intentional attitudes because those attitudes are necessary to understand instances of behavior as actions. What is it that permits us to describe an event as Felix extinguishing the kitchen fire? (Rather than, e.g., Felix doing a dance with the fire extinguisher.) It will involve the fact that Felix wanted the fire to go out and believed that using the extinguisher was a way to make it so. Intentional attitudes are about the objects that feature in Felix's action—the fire and the extinguisher—and if Felix lacked such attitudes, it would be impossible to orient him and his doings within the world, and thus impossible to say that what he did was perform a particular action intentionally.

Intentional attitudes are about things in the world. My belief is about this mug, not that book. My intention is to blow up Parliament, not to paint Windsor Castle. But how do my beliefs, desires, and intentions come to have this "aboutness"? One way to approach this question is to consider how we come to ascribe such attitudes. What leads us to attribute Nigel with an intention to blow up Parliament, rather than one to paint Windsor Castle? The usual ways are familiar enough. We see Nigel planting a bomb, he tells us that he has this intention, we observe him collecting explosives and studying blueprints, we know that Nigel hates Parliament and tries to solve all his problems by blowing things up. When we do these things—observing Nigel's behavior, interpreting his utterances, and considering his other intentional attitudes in an attempt to paint a portrait of Nigel that makes sense of him—we are engaged in an interpretation of Nigel. We have taken up a distinctive explanatory stance, what Dennett calls the "intentional stance". This is an explanatory paradigm whose posits include intentional attitudes, whose evidence is anything publicly available about an individual, whose methods reflect constitutive standards of action, and whose goal is to make sense of that individual and what he does as an agent.

That we actually proceed in understanding people's intentional attitudes in this way seems beyond dispute. But I want to go a step further and say that not only is interpretation our usual method for trying to understand agents and their intentional attitudes, but interpretability is a *condition* on their having those

attitudes. For an agent to have an intentional attitude—like an intention to blow up Parliament—they must be interpretable as having it.[6]

Obviously, some hedging is needed here. There are things about Nigel we may not realize, and this may complicate our interpretation. We may not know that Nigel's hot-headed talk about blowing up Parliament is a feint to distract attention from his plan to dump two tons of fish in Royal Albert Hall or that he is a performance artist engaged in an elaborate project to problematize the Twenty-First Century Terror State. To make this premise minimally plausible, we must understand interpretability in a way that accommodates the fact that actual interpreters can be both ignorant and misled. The natural way to do this is to bracket cases of ignorance like these and tie interpretability to interpreters who are, as Davidson puts it, "fully informed" (2001: 148), interpreters who know about Nigel's membership in an avant-garde arts collective, his late night soliloquys, and his transactions with the fish monger.[7] More generally, we can say that the interpreter knows everything that could be known about a subject by observing them. With this amendment, the thesis in question becomes: interpretability by a fully informed interpreter is a condition on having an intentional attitude.

The claim that all intentional attitudes are interpretable follows from two further claims. The first I call *Wittgenstein's Thesis*. In order for an agent to have an intentional attitude, it must be about these particular objects in the world, and for this to be the case there must be some fact—call it *F*—about the agent's relationship to those objects that fixes this content, and because *F* is about the agent's place in the world, *F* must be public, which is to say in principle available to potential interpreters. The second thesis I call *Davidson's Thesis*. It is the claim that the significance of *F*—that it indicates a particular intentional attitude—is established only within a distinctive explanatory paradigm, the paradigm of interpretation.[8]

I attribute the first thesis to Wittgenstein because his claim from the *Philosophical Investigations* that "an 'inner process' stands in need of outward criteria" is a familiar source of inspiration for publicity constraints on mental states (1953/2009: §580). But what I am claiming is weaker, since I am only interested in intentional attitudes. I could accept that Nigel has a private *sensation* that is undetectable by the methods of even by a fully-informed interpreter. But to say

6. This is a thesis that has been defended, in different forms, by Donald Davidson, Daniel Dennett, and David Lewis. See William Child (1996: 7–22) for a discussion of how to understand it.

7. How to flesh this out is a matter of significant controversy, which I can't do justice to here. For a detailed reading of Davidson's attempts to spell out the standard, see Ernest Lepore and Kirk Ludwig (2005: 156–166).

8. These two premises are similar to arguments (iii) and (ii) respectively that Child entertains for the necessity of interpretability for thought. See Child (1996: 33–37).

that Nigel has a certain belief or intention is to say that he has successfully oriented himself—representationally or deliberatively—about a definite point in the world, and whether he achieved this orientation will depend on facts about the connection between him and that point.

In an extreme case we could imagine Nigel *trying* to intend to blow up Windsor Castle but failing to form such an intention through misidentification. Nigel wants to blow up William the Conqueror's great fortress but his shoddy travel guide mixed up Windsor Castle and Buckingham Palace, and he makes all his plans for the destruction of the latter. When we say that Nigel has formed an intention to blow up Windsor Castle, we are crediting him with a kind of success that precludes this kind of possibility. He has "gotten on" to a particular part of the world. Thus intention and belief are success states. To say that someone has a belief or intention with a particular content is to credit them with the achievement of "getting on" to certain parts of the world.

What is it that makes Nigel's intention about Windsor Castle—and not Buckingham Palace, Big Ben, an ornery sea otter, the mound of earth directly underneath Windsor Castle, undetached castle parts, or something else? Some of the answer will involve Nigel's other intentional attitudes. But adducing these further attitudes only carries us so far. For we can ask how this whole web of attitudes manages to "get on" to one coordinate set of objects in the world rather than another. And so eventually this line of questioning must bring us to Nigel's interface with the world—for example, his behavior and what is available to his senses—and thus to features of Nigel that are publicly available.

To dramatize the point, imagine that none of Nigel's behavior, nor indeed anything publicly available about him indicates an intentional connection between him and the event of Windsor Castle exploding. It seems difficult to imagine crediting Nigel with successfully "getting on" to this event, since we cannot explain why it is an intention about Windsor Castle rather than Buckingham Palace, Big Ben, or an ornery sea otter. We might have been tempted to think that Nigel's first-person perspective on his intention suffices—that somehow his *taking it* to be an intention about Windsor Castle is enough. But that will not do, as the mistaken travel guide example shows. Without some connection between Nigel and the event in question we cannot credit him with successfully having an intention *about* that particular event.

To summarize, the thought behind Wittgenstein's Thesis is that having an intentional attitude is not something an agent can do unilaterally, since it involves a kind of success in latching onto part of the world. And whatever this success comes to, it will be partly constituted by some facts *F* that are themselves in the world and so publicly available to interpreters.

The second step of the argument for the interpretability of all intentional attitudes is to show that given (a) the existence of such *F* and (b) the fact that

Nigel really does have an intention to blow up Windsor Castle, it follows that Nigel will necessarily be interpretable as having this intention. This follows from Davidson's Thesis: it is *only* by figuring in a distinctive mode of explanation—interpretation—that *F* gains significance as contributing to a particular intentional content. That is, it is *only* as part of an interpretation of Nigel that *F* partially constitutes Nigel's having an intention to blow up Windsor Castle.

Talk of thoughts, Davidson says, "belongs to a familiar *mode of explanation* of human behavior and must be considered an organized department of common sense which may as well be called a theory" (2001: 158, my emphasis). Dennett says something similar. For him interpretation is a kind of explanation shaped by a distinctive stance that we can take toward a system. He calls this the intentional stance, and contrasts it to the stance we take when we try to explain events in causal-mechanical terms (the physical stance) or as working out the logic of some design (the design stance). Taking up the intentional stance "consists of treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental states exhibiting what Brentano and others call *intentionality*" and supposing that those states and the object's outward behavior are connected by principles of constitutive rationality (Dennett 1989: 15).

I think Davidson's Thesis is most plausible when understood as an instance of a more general claim. It is sensible to posit theoretical entities only relative to a distinctive "mode of explanation" that specifies the conditions for and individuation of such posits. For example, if we are trying to understand an unusual creature as a mammal, and are specifically trying to identify its organs, we will encounter reasons for and against classifying one of these organs as a uterus. But obviously the positing and identification of a uterus is tied to what we might grandiosely call the mammalian mode of explanation. It makes no sense to talk about certain facts about the animal as qualifying this thing as a uterus independent of a larger scheme of organ individuation deployed when trying to understand something as a mammal. The classification of something as a uterus makes no sense independent of this background scheme. By the same token, Davidson's Thesis maintains that it makes no sense to talk about a fact *F* partially constituting Nigel's intention as one of blowing up Windsor Castle outside of the context of the distinctive mode of explanation in which we attribute intentional attitudes—the method of interpretation.

This claim is borne out by our practices. Suppose you attribute an intention to blow up Windsor Castle to Nigel and cite the fact that he planted bombs around the perimeter of Windsor Castle as evidence. I might challenge this ascription by adducing further facts about Nigel. I could say that the bombs are filled with paint and Nigel is a guerilla artist. Or that Nigel believes that the ground beneath Windsor Castle is haunted and he wants to drive out the spirits. Or that Nigel is a devoted fan of the New York Mets. The way to decide which of

these facts matter to our interpretation of Nigel—and how—is by trying to produce an interpretation of him conditioned on each of them and seeing what happens. Given that Nigel planted some bombs and is a guerilla artist, is it still our best interpretation of him that he intends to blow up Windsor Castle? Given that Nigel planted bombs and loves the New York Mets, what follows *then*? What we see in these examples is that the relevance of some fact to Nigel's intentions is not an intrinsic feature of that fact, but something that emerges only as part of a comprehensive interpretive explanation. That's what Davidson's Thesis says.

The combination of Wittgenstein's Thesis and Davidson's Thesis yields our second premise. By Wittgenstein's Thesis there must be some public facts *F* that constitute an agent's successfully possessing an intentional attitude. And by Davidson's Thesis, the existence of *F* guarantees the interpretability of the agent for the simple reason that it is only as part of an interpretive explanation that this constitution is achieved. And this gives us the claim that all intentional attitudes are interpretable.

With this conclusion in hand, let's return to the main line of argument. The first two premises of the argument entail a claim I will call a lemma:

Lemma 1. Interpretability is a condition of agency.

As I have said, this lemma only follows if we understand "interpretable" liberally. The lemma does not say that I cannot act without that act being successfully interpreted by other agents. But the lemma is not without teeth. A person can act in ways so strange, incoherent, or self-stultifying that they are uninterpretable, even on arbitrary improvements of the interpreter's epistemic position. According to the lemma, this person cannot be credited with the usual battery of beliefs, desires, and intentions, and so cannot be understood as performing actions that depend on her having those attitudes. And that makes them something less than an agent.

## 5. Reflective Agency

Lemma 1 is an interpersonal condition on agency, the kind of condition that I said Korsgaard and Velleman should be interested in offering but don't. So are we done? Have we planted morality in the soil agency? Not quite: showing that interpretability is a condition on agency does not entail that a creature must *regard* or *treat* his interpreters (or anyone else) in any particular way in order to qualify as an agent. All he need do is carry on in the right ways, and nothing I have said so far suggests that these ways need to involve anything distinctively moral. So we need to say more.

If we can show that full-fledged agency requires some kind of *sensitivity* to the conditions of interpretation, and not just conformity to them, then this may allow us to argue that an agent must evince a certain regard for her interpreters. This is what I shall try to do, but it requires narrowing our focus.

Distinguish two kinds of agency. Animal agency consists in an ability to act on the basis of one's intentional attitudes in whatever the distinctive way required for genuine action is. Reflective agency involves animal agency plus successful reflection on the attitudes that produce action. That is, the reflective agent is one who steps back from her beliefs, desires, and intentions to scrutinize them, to examine how they hang together, to ask whether they amount to *good reasons* for action. The reflective agent only acts on grounds that survive this scrutiny. (One might argue that there is no such thing as animal agency since the possession of intentional attitudes requires reflection. I'm sympathetic to this idea, but will not assume it.)

This distinction is a cousin of one in epistemology. As Sosa explains it, animal knowledge "requires only that one track reality", whereas reflective knowledge also "require[s] broad coherence, including one's ability to place one's first-level knowledge in epistemic perspective" (1997: 422, 427).[9] Similarly, animal agency requires only that one be efficacious in asserting one's will, whereas reflective agency requires that you undertake adequate levels of self-scrutiny and maintain a minimum level of self-imposed harmony. In what follows, I will shift from talking about agency as such to talking about reflective agency. My claim is that *reflective* agency requires sensitivity—not just conformity—to the conditions of interpretation outlined above.

This shift prompts two concerns. The first is about what reflective agency is. What kind of reflection is required? How much is required? At what point does reflection get in the way of other demands of agency? Over what intervals can we assess reflective agency? For the most part I would like to leave these questions aside and say that my argument is compatible with many different answers to them. The second concern is more pressing. Can reflective agency anchor a constitutivist argument? After all, the whole conceit of constitutivism is that one's own agency is a commitment generic enough that no one can relinquish it. But if reflective agency is just one kind of agency, can't it be rejected in favor of its animal alternative?

Constitutivists are wont to say that agency is a unique source of normative requirements because it is "inescapable": if I act, I cannot help but be an agent. Of course, agency *is* escapable in a certain sense. I can die. I can put myself into a coma. I can violate the constitutive requirements of agency. So the "inescapability" of agency that matters for constitutivists must be of a different sort. The

9. We could also, slightly tendentiously, call this "human agency" as Charles Taylor (1985)

right way to think about the idea, I propose, is that agency is a necessary *presupposition* of the deliberative standpoint and thus of the standpoint from which we entertain practical questions. If we are asking practical questions about what to do, or normative questions about what our reasons are, we are already assuming that we are agents. In this way, the constitutive demands of agency function as background principles for all deliberation. It is this status that gives them their special normative authority.[10]

Notice that the demands of reflective agency enjoy the same status, for it is not as agents *simpliciter*, but as reflective agents that we entertain normative questions. Asking whether some consideration really does constitute a reason to act in such and such a way, questioning whether the object of some desire is really good after all, wondering whether this feature of ourselves makes it right to behave in this way—these questions are exercises of *reflective* agency. Because the animal agent does not reflect on the bases for which she acts, she never asks these sorts of questions. Thus it is not merely agency in general that we presuppose in normative thinking, but reflective agency in particular.

I cannot litigate the constitutivist's argument from all angles here. My point is just that on the best understanding of that argument, reflective agency makes just as good of a normative foundation as agency *simpliciter*. If we are inclined to think that norms can indeed be grounded in the constitutive demands of agency, then we should also believe they can be grounded in the constitutive demands of reflective agency, since it's this latter sort of agency that is presupposed by the entertaining of normative questions.

This is my third premise, the only explicitly constitutivist premise in the argument: the constitutive requirements of reflective agency have universal normative authority. Because the entertaining of normative questions presupposes this capacity, everyone who entertains such questions is bound by these norms. In the next section I look at the results of combining this premise with our first lemma.

## 6. The Demands of Reflective Agency

Let's examine the conditions of reflective agency in light of Lemma 1. Reflective agency involves reflecting on the reasons one has to perform certain actions and being sensitive to considerations favoring or opposing those actions in a systematic way. Here I am interested in an important subset of these considerations: facts about the *conditions* on their performance. If I *cannot* perform a given action, then adequate reflection should lead me to rejigger whatever complex of atti-

10. Different ways of cashing out this point can be found in Luca Ferrero (2009), Matthew Silverstein (2014), and Walden (2012).

tudes might lead me to try to perform it. For example, if a man attends a baseball game with the intention of umpiring from the bleachers, if he persistently tries to buy things without any money, if he tries to knight his beer buddies, if he plans to take my rook in a game of checkers, if he habitually sets about to eat quantities of food of greater than his own mass, if he plans to lift the moon with a garden trowel, then we would be right to say that something has gone wrong with his reflective agency. For he has failed to apply one important kind of scrutiny in his practical reflection: the regulation of one's actions in light of the conditions on the possibility of their performance.

This my fourth premise: reflective agency requires sensitivity to the conditions on the actions that one may entertain performing. Thus agents must be aware of and, to some extent, guided by the conditions on their performance of a candidate action. Combining this with Lemma 1, which says that interpretability is a condition on agency, we get:

> Lemma 2. Reflective agency requires sensitivity to the conditions of interpretability.

I have chosen the vague word "sensitivity" advisedly, with the goal of leaving open, at least for the moment, what it involves. I obviously do not mean to suggest that one must be constantly reflecting on their interpretability in order to act, nor that the considerations they discover in these reflections produce overriding or even particularly strong reasons. What, exactly, the best habits of reflection are is an important question, but not one we must answer now.

Instead I want to ask two more general questions: *To what (or whom)* must agents be sensitive? And *what form* must this sensitivity take?

## 7. Sensitivity to What? And How?

In my list of impossible actions above, the conditions relevant to the impossibility fall into two categories. Some of them are what we might call *brute*. It is a fact about a man's gut that he cannot eat more than his own weight in hotdogs, about the weight of the moon that it cannot be lifted with a garden trowel. Others are *institutional*. These are closely related to Austin's "felicity" conditions (1962: 22ff.) and Searle's "preparatory rules" (1969: 64ff.). I cannot fire a man unless I am his boss. I cannot sentence a criminal unless I am a judge. I cannot strike out a batter unless I am a pitcher. These conditions reflect not brute facts about guts and celestial bodies, but about institutions of employment, law, and baseball.[11]

11. The distinction between brute and institutional is also Searle's (1969: 50ff.).

Which of these two classes better represents the conditions of interpretability will matter a great deal to how we answer the questions I ended the previous section with. If the conditions on agency related to interpretability are brute conditions, like my inability to eat more than my own weight in hotdogs, then the sensitivity I need to display toward them will take one form: it will involve something like being aware of simple, non-negotiable facts about what is interpretable and what is not, and using them as I would use a roadmap or blueprint. But if they are institutional—if they are like the conditions on my performing the illocutionary act of marrying two people—then they will take a rather different form. In this case, one must be sensitive not only to facts, but also to other people who exercise authority by virtue of the relevant institution. Thus the batter needs to be sensitive to the umpire whose verdictive judgment determines whether he is out on strikes. This authoritative party may be an individual, like a judge or umpire, or it may be a small group (a legislature) who make laws, or, arguably, it could even be the whole moral community that endows me with the power to make promises. In these circumstances the appropriate sensitivity will not just involve consultation (as with a roadmap or blueprint), but recognition of the authority of this party. Even if he makes all the right motions, a man cannot really be said to be playing baseball if he doesn't recognize the umpire's special role in calling players out. Recognition in this sense is a very different attitude from a player's awareness that his injured hamstring makes it impossible for him to catch the ball.

We should be interested, then, in whether the conditions of interpretability are brute or institutional. There is a straightforward, *prima facie* reason to favor the view that interpretability is an institutional condition. Interpretation is a practice, one dependent on other practices: social and cultural norms, linguistic conventions, and the "dramaturgical" principles of face-to-face encounters (Goffman 1959). And the conditions of success within a practice are, by definition, institutional.

To this one could object that physics is a practice, and yet few would say that the conditions on its posits—being an electron—are institutional. There is an important difference between these two cases, though. Our notion our interpretability is fixed by the practice in a way that "electron" is not fixed by practice of physics. We have no concept of "interpretability" except as success in the practice of interpretation. By contrast "electron" doesn't just denote a particular theoretical outcome in the practice of physics; it points to something beyond that practice. This is true even for the idealized notion of interpretablitiy we are working with. When we say someone is interpretable because an "ideal interpreter" could offer an interpretation of her if sufficiently apprised of her behavior and utterances, we are not imagining an omniscient god opening her mental box and sneaking a peek at her beetle. Whether or not this kind of epistemic access

is even conceivable, it is certainly not an interpretation. What we are imagining is someone more or less like ourselves engaged in more or less the same interpretive practice we are, albeit with more evidence and cognitive resources. We are not imagining a radically different kind of discovery—the uncovering some fact wholly independent of this practice. Thus our concept of interpretability is bound to the practice of interpretation in a way that distinguishes it from the concepts of natural science.

Another objection to the claim that interpretability is an institutional condition begins with the observation that there a significant difference between the power of an individual interpreter and the power of a judge or umpire in stereotypical institutional cases. If an umpire calls me out on strikes, he *thereby* renders me out. An actual person's power over my intentional attitudes is much less direct and decisive. His inability to interpret me does not necessarily mean I lack an intentional attitude or am not an agent. Maybe I'm misleading him. Maybe he is missing vital information. Maybe a mistaken assumption has led him astray. This is why we must insist that the relevant condition on agency is not interpretability by any actual person, but by a sufficiently well-placed interpreter. That interpretability involves such idealization marks a significant difference between it and our paradigms of institutional conditions.

We shouldn't exaggerate this difference. Umpires' calls, judges' verdicts, and legislatures' laws are the most-cited examples of individuals creating conditions on our ability to perform certain actions precisely because those conditions are so clear-cut. But there are many cases where this conditioning is much messier: cases where the conditioning is itself conditional, where it is merely *pro tanto*, and where it is defeasible. An individual legislator's vote can partially constitute the law that makes it possible for me to get married, but it does so only conditional on enough other legislators voting the same way. A judge's verdict may make it impossible for me to vote, but this verdict may be vacated if it is discovered that he was bribed. Riff's low opinion of me may keep me from rolling with the Jets, but the opinions of the other members of the gang matter too, so the force of Riff's is merely *pro tanto*. These are all examples of institutional conditions on actions, and yet in each of them the relationship between an individual authority and those conditions is more complicated than we see in the example of balls and strikes.[12]

Can we imagine a practice whose relationship to some status is "messy" in roughly the same way that the practice of interpretation's relationship to agency is messy *and yet* whose conditions are plainly institutional? I think we can. Suppose a group of friends and I are engaged in a game of make-believe. The rules of this game are fluid and tacitly negotiated by its players, but there definitely are

12. Austin himself comments on this (1962: 151ff.).

rules. I can't pretend to be drinking tea by giving you a noogie, and I can't pretend to go to the moon by taking off my trousers. What kind of behavior counts as what action within the pretense depends, more or less, on the judgments of other players. I say "more or less" because we all have various limitations and idiosyncrasies that lead us to occasionally interpret and apply the standards of the game in distorted ways. Olaf is a hothead and too ready to interpret behavior as hostile. Felix has a prurient streak and sees obscenity in everything. To keep the game from spinning out of control, we try to bracket off these eccentricities of interpretation. Just because Olaf thinks you are challenging him to a pretend fight doesn't mean you are. The rules also leave room for error in another direction. One player may prematurely interpret another without realizing he was performing a temporally extended action: Felix may think that I am pretending to dig a grave, when in fact I am enacting a much more elaborate pretense about excavating a pharaoh's tomb. The rules permit us to say Felix to be mistaken here, and for the players to discover this when I pretend to drag up a sarcophagus.[13]

The rules of this game are player-dependent but idealized. What counts as pretending to $\phi$ is not *just* whatever the gang says, but what the gang *would* say if they weren't so hotheaded or prurient or they waited just a moment. In this way, the relationship between interpretation and agency is structurally similar to the relationship between the narrow form of interpretation that governs this game and the particular class of actions pretending-to-$\phi$. It goes without saying that the former relationship is more complicated, but the make-believe game does involve much the same "messiness" that gave us pause about understanding interpretability as completely fixed by the practice of interpretation: the idealization, the defeasibility, the distribution of the conditions amongst many individuals. Now, it would obviously be a mistake to say the conditions of pretending to $\phi$ are brute rather than institutional just because these conditions involve significant idealization. It would be a similar mistake, I suggest, to say the conditions of interpretability are brute.

Let's agree, then, that the conditions of interpretability are institutional and not brute. What does this mean for our two questions: *to what or whom* must we be sensitive in order to qualify as reflective agents? And *how* must we evince this sensitivity?

The analogy with the make-believe game is instructive for both questions. To participate in the game of make-believe I must be sensitive to the other players and their ability to make sense of me. Felix's understanding may be distorted by his prurience, but that doesn't mean that I can play the game while ignoring what he makes of my pretense. It is likewise a condition on our reflective agency that we are sensitive to other interpreters—*even if* those other interpreters may

13. Thanks to John Kulvicki for suggesting games of make-believe to me.

be epistemically poorly placed to offer a good interpretation of me and so the effective force of their judgments is tempered in the same way that our make-believers' are. Sometimes this tempering will be local. I may be deliberately trying to deceive you. Other times it will be systemic. I may be an avant-garde performance artist whose intentions are accessible only to the most astute critics or a member of a clique with self-consciously unconventional manners.[14] In both sorts of case, the confusion or misapprehension of (most) others should have negligible effect on my standing as an agent. But just as I must maintain some *principled* sensitivity to Felix *qua* interpreter in the make-believe game despite his shortcomings—an openness to the significance of his interpretations, even if they are commonly based on misapprehension—I must do the same for poor critics of my performance art and those on the outside of my clique. I must do this for two reasons. First, the ignorance that makes them poorly situated interpreters is a contingent condition, so they are, in principle, part of my interpretive audience. Second, and more importantly, my performance art and fancy manners are situated in and depend on larger, more generic practices of action and interpretation. Even my poorest critics are constituent members of these practices. It is the institutionality of interpretability conditions and the standing that these individuals have within the relevant institution that ultimately requires some minimal and basic sensitivity to their understanding of me.

We observed that for me to be sensitive to the institutional conditions of baseball, I need to be sensitive to umpires, and this kind of sensitivity is importantly different from the kind I display when knowledge of my injured hamstring keeps me from diving for a ball. It is a sensitivity to another person's standing to make judgments, render verdicts, and advance claims. It is a *recognition* of their authority. The sensitivity we must evince as a condition on reflective agency will likewise take the form of the recognition of institutional authority. We must recognize other persons *qua* interpreters as having the authority to interpret us in a way that is in principle relevant to our success as agents. We must do this for the same reason we must recognize the analogous authority of umpires and judges—because they occupy an office designated by the practice as having that power. This will not be arbitrary or absolute authority, of course. Interpreters cannot refuse to make sense of me on a whim, and their interpretations are governed by rules. And if they are poorly positioned to understand me for whatever reason, then their authority is significantly curtailed. But interpreters are no different from other authorities in these respects.

The authority possessed by interpreters is what Stephen Darwall (2006) calls "second-personal". He explains this kind of authority with an example. You are stepping on my foot and I exclaim, "Ouch, stop it!" We can understand the au-

14. Or speak in an unbreakable code, as Davidson (1994: 121) suggests.

thority of my command and the reason it provides you to release my foot in two ways. It may be that the reason you have is one grounded in the general disposition of the world. Your stepping on my foot causes unnecessary pain to exist, and you have a reason to extinguish it. In this case the authority that I have is epistemic. I am an expert on pain in my little corner of the world, and so if I tell you that you are creating unnecessary pain, you should listen to me because I am giving you reliable testimony about the world that will give you reasons to act. On the other hand, we can think that the reason you have to release my foot does not arise from the fact that doing so would decrease the amount of pain in the world, but from the fact that I have *objected to* your stepping on my foot. In this case my command is not an indicator of your having a reason to release my foot, but constitutive of it. Here the authority I have is not the authority of someone with sure-footed access to the reasons the world provides, but that of someone with the *standing* to create reasons for you by making claims. These reasons are similarly second-personal insofar as they are "grounded in the (*de jure*) authority relations that an addresser takes to hold between him and his addressee" (Darwall 2006: 4). If the conditions of interpretability were brute, then the best other interpreters could do would be to know brute facts about interpretability, in the way someone could know about the badness constituted by the pain in their foot. This person would not have second-personal authority, but epistemic reliability. Because the conditions of interpretability are institutional, however, interpreters have authority *not* by dint of their access to some further fact "beyond" the practice of interpretation. They possess their authority in virtue of occupying a particular office.

Second-personal authority is the authority to address claims, demands, and expectations. The interpreter's standing empowers her to address claims, demands, and expectations of a particular species, what I will call claims of intelligibility. "You plan to join the Foreign Legion? What a senseless thing for someone like you to do." "You believe that fluoride robs us of our precious bodily fluids? That's senseless." "You couldn't *really* want to marry your dog. I can't even fathom your having that desire." "Given the sort of man you are, it'd make much more sense for you to go to Harvard than Princeton." "Since you burned down that church, you haven't been the same person." These proclamations may look like epistemic reports, but in context they perform distinctive interpersonal functions: giving you advice, expressing and enforcing standards, or holding you accountable. In this context such utterances are not mere reports but claims because—to use Darwall's language—they come with an RSVP. They are made with the expectation that you will receive them and respond to their force, either by acceding to them or rebutting them. When I say that it'd be unintelligible for you join the Foreign Legion or senseless to believe that fluoride robs us of our precious bodily fluids, I am

addressing a normative *expectation* that you will not do these things. (This isn't to say that there is no epistemic aspect to these claims—that I can make claims of intelligibility unmoored from the evidence. Obviously, I cannot. But in this respect claims of interpretability are no different from those made by a judge or umpire.)

In this section I have argued that the sensitivity agents must have to the conditions of agency takes the form of recognizing the second-personal authority of interpreters. I must recognize Felix's authority to address interpretative claims to me. This is true even if Felix is in fact in a bad spot to interpret me (because I'm trying to deceive him, or he doesn't know the first thing about my religious rituals, or . . .). The reason I must do this is that Felix is one constituent part of the practice of interpretation that conditions my possession of intentional attitudes. This endows Felix with a *de jure* second-personal authority, even if the *de facto* force of his interpretative claims is very "messy" in the ways canvassed above—that is, even if the actual connection between his interpretation and my agency is *pro tanto* and defeasible.

## 8. Summarizing the Argument

I began by saying that the holy grail of constitutivism was establishing the universal and unconditional authority of morality in the conditions of agency. I was pessimistic about extant attempts to do this because they focused on features of agency that were too individualistic to lend themselves to grounding other-regarding principles. I have now provided an argument that purports to show how other-regarding principles can be so grounded. It goes like this:

For a creature to be an agent, she must have intentional attitudes.
For a creature to possess an intentional attitude, she must be interpretable as having that attitude.

Lemma 1. Interpretability is a condition on agency.

The constitutive requirements of reflective agency have universal normative authority.
Reflective agency requires sensitivity to the performance conditions of the agent's prospective actions.

Lemma 2. Reflective agency requires sensitivity to the conditions of interpretability.

Sensitivity to the conditions of interpretability requires the recognition of the second-personal authority of interpreters to make claims of intelligibility.
Recognition of the second-personal authority of interpreters to make claims of intelligibility is a condition on agency.

From this we can conclude:

R. It is a universally authoritative requirement that one recognize the second-personal authority of interpreters to make claims of interpretability.

(R) is an other-regarding principle. Establishing such a principle is half the battle for a constitutivist hoping to vindicate morality. But it's obviously not a total victory, since we might still wonder what it has to do with paradigmatically moral sorts of recognition. I will try to spell out this connection in the next section.

## 9. Respect for Humanity

It should be clear that (R) prescribes some manner of respect. It requires agents to acknowledge the standing of others by regarding certain claims made by them as possessing unconditional and non-strategic practical significance. This is the essence of respect, at least according to some standard analyses. For Kant respect involves regarding someone not as a means to some further end, but as an end in themselves, namely as someone *for whose sake* I act.[15] For Darwall (2006: 119ff.) it is the recognition of the standing to address second-personal claims and the reasons that follow from them. Sarah Buss (1999) describes respect as experiencing others *as subjects*: respecting a person means experiencing oneself as an object for them—what Sartre calls *être-pour-autrui*—and in so doing briefly taking on their practical point of view as our own.[16] (R) requires all these things. I must recognize other persons, *qua* interpreters, as having the authority to address claims, which can in turn give me second-personal reasons. When I act on these reasons I am regarding them not as a means to some further end, but as something for whose sake I act in a particular way—as an end in themselves. In paying this recognition, I necessarily see myself as an object for this other person—as an object of interpretation—and in so doing I take up *their* perspective.

So (R) entails *something* in the way of respect. But on first inspection, it seems to be a rather anemic sort. According to (R) we must respect others *qua* interpret-

15. *Groundwork for the Metaphysics of Morals*, 4:428.
16. See Sartre (1943/1956: Part III, Chapter 1).

ers, but not necessarily beyond that office. This narrowness is easily dramatized. I pistol-whip Felix in a dark alley in a bad part of town and demand his money. He objects, but I insist that I will murder him if I am not given satisfaction. In this scenario it appears I am living up to (R), since mugging a man in a dark alley is a perfectly intelligible thing to do in the bad part of town. (We could even stipulate that if Felix is confused by my behavior I will take pains to make it intelligible to him.) And yet I am also quite obviously not respecting Felix in the usual, moral sense, since I am ignoring his reasonable objection against being robbed.

I want to suggest that this appearance is misleading. In fact (R) commits us to a full-blooded respect for persons. To see why, we need to consider more carefully *what it is* we recognize in another person when we recognize their authority *qua* interpreter. This will obviously depend on which powers and capacities are essential to the endeavor of interpretation. On this is question, I propose to follow Dennett. He describes interpretation thus:

> You decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (Dennett 1989: 17)

The crucial idea here is that interpretation is not undertaken by a *sui generis* Interpretive Faculty—by a mental module whose sole function is to issue interpretations. Rather, it involves the vicarious deployment of our generic capacities for theoretical and practical reasoning. We entertain hypotheses about a subject's motives, input them into our own reasoning, and then compare the outputs with what we have observed. Davidson and Lewis's respective versions of the principle of charity commit them to the same model. To say what it would be charitable for a person to believe or desire, we must pass a judgment about what one *ought* to believe or desire, and this kind of judgment requires the use of our own rational faculties. Interpretation is a form of rationalizing explanation, so rendering an interpretation means having a sense of what it would be sensible for a person to do and believe, and that means passing judgment with our own faculty of reason.

If this is right, then it seems that all that is strictly necessary for holding the office of interpreter is being capable of reasoning. But then anyone who possesses these capacities will fall within the scope of (R). Moreover, this power of reasoning seems to be *what it is* that we recognize when we recognize others

*qua* interpreters. When I acknowledge Felix's authority *qua* interpreter, I am acknowledging the authority whatever it is that enables him to perform this office, which is his "rational nature"—his ability to reason along with me as I confront practical problems.

But this capacity is none other than what Kant calls as our "humanity": the capacity to set ends through reason, and more generally to determine our will in accordance with rational reflection.[17] Thus in recognizing other persons *qua* interpreters *what it is* that I am respecting is their humanity. If we supplement (R) with this understanding of interpretation, we get the conclusion that we are in fact required to recognize the authority of humanity in all who possess it. This yields:

> S. It is a universally authoritative requirement that one recognize the second-personal authority of humanity in all persons.[18]

(S) looks like a step in the right direction, appearing, as it does, rather closer to the Formula of Humanity than (R). Nonetheless, we might worry that (S) is consistent with a highly circumscribed form of respect. Return to the mugging case. Even if, per (S), what it is I am required to respect in Felix is his humanity, it's not clear that I am required to recognize *all* the claims that might issue from that capacity. Our argument might show that I am required to recognize the authority of his humanity to make what I have called "claims of intelligibility", but *not* to recognize the practical significance of his objections against being mugged. In other words, we might worry that we are entitled to:

> T. It is a universally authoritative requirement that one recognize the second-personal authority of humanity in all persons to make *claims of intelligibility* (but not others).

But the stronger principle, the moral principle is:

> U. It is a universally authoritative requirement that one recognize the

17. See *Groundwork* 4:437 and *Religion within the Boundaries of Mere Reason*, 6:26.

18. This makes the class of creatures meriting respect large, but perhaps not as large as we would like. (R) does not entail anything about our obligations to creatures lacking humanity. *Prima facie*, this would seem to include most if not all non-human animals, young children, and adults with cognitive disabilities. In this respect, (R) leaves us in the same place as the Formula of Humanity. It is a familiar objection to the Formula of Humanity and Kantian ethics in general that they exclude non-rational creatures from the moral community in this way. I am sympathetic to the idea that some form of moral significance can be secured for non-rational creatures within Kant's system despite this appearance, and I think that this sort of story can be adapted to my framework. But it's too long story a story to tell properly here.

second-personal authority of humanity in all persons in *all claims it might make.*

Are we entitled to (U) or just (T)? The possibility we need to consider in answering this question is one of selective respect for humanity. Can I coherently respect your humanity insofar as it makes claims of intelligibility but not insofar as it makes claims about right and wrong, good and bad, fair and unfair?

I am inclined to say that I cannot. My reasons are related to the holism of normative judgments. Felix's intelligibility claims do not constitute a closed and autonomous system of practical judgments. They are woven into a tapestry of other normative judgments: ones about what is good and bad, virtuous and vicious, right and wrong. Insofar as Felix is rational, this web will be governed by standards of coherence and regulated by the ideal of a harmonious practical outlook. This means that Felix's normative judgments depend on each other. His normative judgments about the wrong of mugging should affect what he thinks about the wrong of blackmail. His normative judgments about when lying is forbidden will affect his judgments about the when we are required to keep our promises. Of course, the dependence of an intelligibility judgment on a moral judgment will be weaker than that between two moral judgments. The wrong of my mugging Felix does not *ipso facto* make it unintelligible. Nonetheless, insofar as moral judgments and judgments of intelligibility are part of a holistic web of mutually supporting normative judgments, this dependence will not be nil: witness the difficulties in trying to make intelligible deep and fundamental moral disagreements and the problem of radical evil.[19]

These coherence constraints governing this web of judgments form part of our conception of rationality. To be a rational creature, one must surpass a minimal threshold of coherence and be minimally engaged in activities that maintain that coherence. Insofar as Felix is a rational creature, the sort of creature who merits recognition, he must maintain a comprehensive and coherent web of normative judgments. If Felix's only judgments of practical reason were intelligibility claims, or if these judgments were completely sequestered from his other normative judgments, then he wouldn't be a fit interpreter and so wouldn't merit recognition. On the contrary, Felix's authority depends on his being a *comprehensive* reasoner: a person who makes all manner of normative judgments and maintains them according to basic standards of coherence.

What does this mean for my respect for Felix? Insofar as I am respecting him *as* a rational creature, it must be this *comprehensive* reasoner that I am respecting. I must respect him as someone who makes a multitude of mutually supportive normative judgments. But then my recognition of Felix's claims must

19. On the former see Philippa Foot (2003). On the latter see Allen W. Wood (2010).

be similarly comprehensive. I cannot respect him as a comprehensive reasoner without recognizing his authority to make a comprehensive range of normative judgments, including normative judgments that function as claims on me: that in mugging him I am wronging him.

This allows us to diagnose the problem with the person who maintains an attitude of selective respect. In effect, this person presumes that the claims of intelligibility a person makes are grounded in a *sui generis* Interpretive Faculty, and that *this* faculty and *only* this faculty possesses the second-personal authority to make claims. But this is a mistake: interpretation cannot be extricated from the entire person, and so the demands of interpretation bind persons together, not mere interpreters.

To put the point succinctly: the claims of humanity form a unified and integrated whole, and so respect for one any part of this whole entails respect for the rest. This I why (S) entails (U).

In this argument the interpretive conditions of reflective agency work like the thin edge of a wedge. According to (R), every agent is required to recognize the second-personal authority of interpreters to make claims of intelligibility on them. But if Dennett's view of interpretation is correct, then the power that qualifies a person for the office of interpreter is none other than their humanity. And the holism of normative judgments means that we cannot consistently recognize the authority of humanity in some of its claims but not others. The recognition required by (R) amounts to a comprehensive and unrestricted respect for the humanity of other persons.

I hasten to add that the "cash value" of this respect remains to be worked out. For all I have said, it could be that Felix's objections give me second-personal reasons not to mug him, but that these are decisively defeated by reasons I have to carry on with the mugging. It could also be that his objections decisively forbid the mugging. These questions will have to be settled in a systematic balancing of claims according to whatever contractualist scheme fits best. I am only suggesting that insofar as we live up to the requirements enshrined in (R), all of Felix's objections to being mugged must be granted non-strategic significance in our practical reasoning—that they are "public" reasons in Korsgaard's sense—not that any particular requirements or prohibitions necessarily emerge from that reasoning.

## 10. Conclusion

In this paper I have criticized two constitutivist strategies for vindicating morality and offered an alternative. In closing I want to compare my proposal with the former and explain where I think it improves on them.

One important difference is in the initial claim about the nature of action and agency. Korsgaard and Velleman suggest constitutive aims of action—self-constitution and self-understanding, respectively—whereas I have suggested a constitutive condition on agency. I don't think that all our actions aim, either explicitly or implicitly, at making ourselves interpretable to others. Instead, I have argued that a kind of sensitivity to the interpretative demands of others is a condition on agency. One advantage of my approach, I suggest, is that this is a better way of developing the basic constitutivist idea, since the idea that all actions share a single aim is one that many find highly implausible.

This approach also produces a rather different form of normative authority for morality. This is especially clear with Velleman. Both Velleman and I see a connection between intelligibility and agency, and we both hope to use this connection to establish the normative authority of morality in some sense. The crucial difference is where other people figure into our stories. For Velleman, it is *self-intelligibility* that forms part of the constitutive nature of action. Other people are merely useful for enacting scenes that allow us to meet this standard. I suggested that this arrangement, even if it nudges us toward pro-moral practices, makes our relationship with other people fundamentally instrumental in a way that is at odds with core moral platitudes. By contrast, on my view intelligibility to others is what matters in the *first instance*. Being interpretable by others is not merely conducive to a self-understanding necessary for successful agency, but partly constitutive of that agency. This makes a crucial difference. For it bestows an unconditional, genuinely second-personal authority on those persons in virtue of their humanity.

This makes my conclusion rather closer to Korsgaard's than Velleman's. But I think my path to this conclusion is surer than hers. Recall Korsgaard's argument, as I have reconstructed it:

1. To interact with another person I must respect her humanity.
2. I must interact with myself.
3. Any proposed restriction of my interactions to a particular set of selves is ill-formed because these selves are constituted by action.
4. Therefore, I must respect the humanity of everyone.

I made two objections. First, premise 3 is false because there is nothing generally "ill-formed" about restricting one's projects by some end whose construction is part of that project. Second, the inference to the conclusion is invalid because it is the same inference that produces the Sorites paradox.

I think Korsgaard is pushed into this problematic reasoning because the features of agency she focuses on do not essentially involve other people. As a result she cannot defend a crucial second premise in what seems like a more natural argument for her conclusion:

1. To interact with another person I must respect her humanity.
2. I must interact with every other person because doing so is a condition on my agency.
3. Therefore, I must respect the humanity of everyone.

But I think we *can* defend premise (2), or at least something in the vicinity of premise (2). I must "interact" with every other person insofar as I must participate in the shared practice of mutual interpretation that can, in principle, include every other rational creature. By appreciating this relationship between agency and other people we get a much more direct argument for the duty to respect humanity.

## Acknowledgments

For helpful comments and conversation concerning earlier drafts of this paper and related material, I am grateful to Roman Altshuler, James Binkoski, Michael Bukoski, Sarah Buss, Luca Ferrero, Dhananjay Jagannathan, John Kulvicki, Samuel Levey, Kathryn Lindeman, Elijah Millgram, Kate Nolfi, Alice Phillips Walden, David Plunkett, Timothy Rosenkoetter, Christine Thomas, two anonymous referees, and an editor for *Ergo*.

## References

Austin, John L. (1962). *How to Do Things with Words*. Harvard University Press.

Bradley, Francis Herbert (1879). My Station and Its Duties. In *Ethical Studies* (145–192). Oxford University Press.

Buss, Sarah (1999). Respect for Persons. *Canadian Journal of Philosophy*, *29*(4), 517–550. https://doi.org/10.1080/00455091.1999.10715990

Child, William (1996). *Causality, Interpretation, and Mind*. Oxford University Press. https://doi.org/10.1093/0198236255.001.0001

Davidson, Donald (1994). Radical Interpretation Interpreted. *Philosophical Perspectives*, *8*, 121–128. https://doi.org/10.2307/2214166

Davidson, Donald (2001a). A Coherence Theory of Truth and Knowledge. In *Subjective, Intersubjective, Objective* (137–154). Oxford University Press. https://doi.org/10.1093/0198237537.003.0010

Davidson, Donald (2001b). Thought and Talk. In *Essays on Truth and Interpretation* (155–170). Oxford University Press. https://doi.org/10.1093/0199246297.003.0011

Darwall, Stephen (2006). *The Second-Person Standpoint*. Harvard University Press.

Dennett, Daniel (1989). True Believers. In *The Intentional Stance* (13–35). MIT Press.

Ferrero, Luca (2009). Constitutivism and the Inescapability of Agency. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 4, 303–334). Oxford University Press.

Fichte, Johann Gottlieb (1797/2000). *Foundations of Natural Right*. Fred Neuhouser (Ed.) Cambridge University Press.

Foot, Philippa Foot (2003). Moral Beliefs. In *Virtues and Vices* (110–131). Oxford University Press

Gert, Joshua (2002). Korsgaard's Private-Language Argument. *Philosophy and Phenomenological Research*, *64*(2), 303–324. https://doi.org/10.1111/j.1933-1592.2002.tb00003.x

Gilbert, Margaret (2014). Acting Together. In *Joint Commitment* (23–36). Oxford University Press.

Goffman, Erving (1959). *The Presentation of Self in Everyday Life*. Anchor Books.

Hegel, G. W. F. (1807/1977). *Phenomenology of Spirit* (A. V. Miller, Trans.) Oxford University Press.

Kant, Immanuel (1785/2002). *Groundwork for the Metaphysics of Morals*. Allen W. Wood (Ed. and Trans.) Yale University Press.

Kant, Immanuel (1793/1996). *Religion within the Boundaries of Mere Reason*. In Allen W. Wood and George D. Di Giovanni (Eds. and Trans.), *Religion and Rational Theology* (39–216). Cambridge University Press.

Korsgaard, Christine (1996). *The Sources of Normativity*. Cambridge University Press. https://doi.org/10.1017/CBO9780511554476

Korsgaard, Christine (2009). *Self-Constitution: Agency, Identity, and Integrity*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199552795.001.0001

Lepore, Ernest and Kirk Ludwig (2005). *Donald Davidson: Meaning, Truth, Language, and Reality*. Oxford University Press. https://doi.org/10.1093/0199251347.001.0001

Reath, Andrews (2006). Legislating for a Realm of Ends: The Social Dimension of Autonomy. In *Agency and Autonomy in Kant's Moral Theory* (173–195). Oxford University Press. https://doi.org/10.1093/0199288836.003.0007

Searle, John. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Sartre, Jean-Paul (1943/1956). *Being and Nothingness* (Hazel Barnes, Trans.). Pocket Books.

Silverstein, Matthew (2014). The Shmagency Question. *Philosophical Studies*, *172*(5), 1127–1142. https://doi.org/10.1007/s11098-014-0340-x

Sosa, Ernest (1997). Reflective Knowledge in the Best Circles. *Journal of Philosophy*, *94*(8), 410–430. https://doi.org/10.2307/2564607

Southwood, Nicholas (2010). *Contractualism and the Foundations of Morality*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199539659.001.0001

Taylor, Charles (1985). What Is Human Agency? In *Human Agency and Language: Philosophical Papers* (Vol. 1, 15–44). Cambridge University Press.

Velleman, David (2009). *How We Get Along*. Cambridge University Press. https://doi.org/10.1017/CBO9780511808296

Velleman, David (2015). Morality Here and There. In *Foundations for Moral Relativism* (expanded ed., 23–51). OpenBook Publishers. https://doi.org/10.11647/OBP.0086.01

Walden, Kenneth (2012). Laws of Nature, Laws of Freedom, and the Social Construction of Normativity. In Russ Shafer-Landau (Ed.), *Oxford Studies in Metaethics* (Vol. 7, 37–79). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199653492.003.0002

Wallace, R. Jay (2009). The Publicity of Reasons. *Philosophical Perspectives*, *23*, 471–498. https://doi.org/10.1111/j.1520-8583.2009.00180.x

Wittgenstein, Ludwig (1953/2009) *Philosophical Investigations.* (G. E. M. Anscombe, P. M. S. Hacker, and Joachim Schulte, Trans.) Wiley-Blackwell.

Wood, Allen W. (2010). Kant and the Intelligibility of Evil. In Sharon Anderson-Gold and Pablo Muchnik (Eds.), *Kant's Anatomy of Evil* (144–172). Cambridge University Press.