IMPROVISATIONAL AGENCY

Benjamin Bagley

A dissertation to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy.

Chapel Hill
2013

Approved by:

Susan Wolf

Gerald Postema

Ryan Preston-Roedder

Ram Neta

Geoffrey Sayre-McCord

ABSTRACT

Benjamin Bagley: Improvisational Agency
(Under the direction of Susan Wolf)

Consider three puzzles in ethics.

- Love should affirm something special about you, and so be a selective response to certain of your qualities. But it should also appreciate you as a particular individual, not as a member of the class of people with the qualities in question.
- A central species of blame is essentially addressed to the agent of a bad action. Yet while it is not aptly directed toward agents who cannot be reasoned into acting better (who can only be written off), it is not plausibly limited to merely procedural failings, like ignorance, confusion, or weakness of will.
- There seems something right about the incompatibilist thought that real freedom requires a radical ability to create yourself, by making choices undetermined by your given past. But, as Hume and many others have argued, choices must be explained by facts about who you are to be meaningfully attributed to you.

Overwhelmingly, philosophers have responded to each puzzle by rejecting one of the propositions that generate it. This, I argue, is a mistake. It reflects an unduly narrow conception of rational agency. Drawing on the phenomenology of improvisation, I show how it is possible to act according to norms you make up as you go, whose content depends both epistemically and ontologically on the particular actions they govern. This enables new solutions to all three puzzles, each a substantial improvement over its predecessors.

For my family

ACKNOWLEDGMENTS

This dissertation was a lot of fun to write. It would not have been fun, and not just because it would not have been written, without a lot of help and goodwill from a lot of people. It is a pleasure to thank these people here.

Susan Wolf's patience, intelligence, and integrity made this project possible. She put up with lots of very bad writing in the early stages of the project, when I most needed someone to have faith in me; her criticism and advice have directly improved every argument, and virtually every page, in the text; her unstinting lack of prejudice and commitment to her students inspired a trust that freed me to focus exclusively on the development of my ideas. She will always be my intellectual and professional role model.

The rest of my committee—Gerald Postema, Ryan Preston-Roedder, Ram Neta, and Geoff Sayre-McCord—all provided very helpful insight and direction. Thanks, in particular, to Jerry for helping me think about Hegel and jazz; to Ryan for lunch at Top of the Hill and emotional support at key times; to Ram for going above and beyond the call of duty as a DRS seminar leader and placement director to engage with the dissertation at a high level before there was even a question of his joining the committee; and to Geoff for being willing to jump in at the last minute and still offer extremely thoughtful and useful feedback.

Wesley Sauret, Nate Sharadin, and especially my officemate Dan Layman made the Caldwell basement a special place to me. I am further indebted to Dan, Nate, and Wesley for helpful discussion about my work, as I likewise am to members of job talk audiences and search committees at Vassar College, the University of California at Riverside, and Stanford University,

to the participants of the Love and Human Agency Project retreat at Tahoe, and to Kristen Bell, Finnur Dellsén, Luke Elson-Meehan, Andrew Franklin-Hall, Jennifer Kling, John Lawless, Errol Lord, Miroslav Losonsky, Jordan MacKenzie, Kate Nolfi, David Reeve, Samuel Reis-Dennis, John Roberts, Chelsea Rosenthal, Jeffrey Seidman, Larisa Svirsky, and (especially) Vida Yao.

I have dedicated this dissertation to my family, with a full heart. But apart from them my biggest intellectual, practical, and personal debts of all are to Susan Rodriguez, the most splendid of all friends.

TABLE OF CONTENTS

1

How I stopped worrying and learned to love rational indeterminacy

1 Introduction

According to a plausible and influential family of views, autonomy (or at least a central form of it) is a matter of governing yourself according to certain norms. The thought is that certain norms—principles, commitments, or, as I'll put it, values—constitute standards of agency to which we ultimately and essentially aspire. They define the perspective from which we assess our actions, and the attitudes and patterns of practical reasoning from which our actions arise, when we step back and ask ourselves whether we really should act, feel, or reason as we do. In this, they define who we essentially are, our real or deep selves. To the extent to which our agency embodies or tracks the values with which we identify in the right way, then, it is up to us. We are active, rather than passive, with respect it. It is *we* who determine its course, rather than some potentially alien forces working within us, which we could intelligibly disavow.

Call this the *normative view* of autonomy. I am interested in a question that arises within the context of the normative view, which I do not think has received enough attention. What happens when the values with which you identify are *indeterminate*, such that there is no fact of the matter, in principle accessible in deliberation, whether they call for you to perform or refrain from a certain action? Can, under these circumstances, you act autonomously? It may seem that the answer must be "no." If you cannot settle the question whether or not you should do something, all things considered—and no amount of prior reflection, no matter how sensitive or

well-informed, could possibly enable you to do so—it would seem that you could not, in good faith, take whatever you *do* end up doing to embody or track your values. And to the extent that defenders of the normative view have considered this question, or formulated their theories in terms that would entail an answer to it, this is indeed the answer they have given. Below, however, I will argue that it is not the answer they have to give.

This is important, for one thing, because it follows from some fairly modest assumptions about the nature of our values that they really are indeterminate in this way, much of the time. In particular, they are often indeterminate in this way in relation to many of the most significant decisions we are called upon to make, ones that play basic roles in shaping our characters and the courses of our lives. If these decisions cannot qualify as autonomous, they would call into question why autonomy, as conceived by the normative view, should really be as important as it would seem, pretheoretically, to be.

I will begin, therefore, by setting out the normative view in a bit more detail and showing why the contents of what it counts as the values with which we identify can often expected to be indeterminate. I will then consider and reject two ways in which defenders of the normative view might respond to the phenomenon, either by denying that it is a problem, or by claiming that we can resolve indeterminacy within ourselves, when it occurs, through primitive acts of will. Finally I will offer my own solution, which is to conceive the contents of the relevant values as things we are essentially in an ongoing process of determining. This, I will argue, not only saves the normative view from having to count certain fundamental, formative choices as non-autonomous, but illuminates a central aspect of their significance.

2 Indeterminacy, ambivalence, and autonomy

It will help to start with an example. The canonical one is Sartre's, of a student who must choose between staying with his mother, who needs him, and leaving her to fight with the Free French (and thereby, perhaps, avenge his brother). Sartre describes this student as *anguished*. He is torn between competing systems of value, and more basically—it is natural to suppose—between competing conceptions of his identity, or the description under which he values himself. Does his relationship to his mother give him reasons to stay with her that override the value of the contributions he might make if he leaves? Is the honorable thing—in the sense of "honorable" that really matters—to care for someone who loves and relies on him, or to resist an oppressor? Would he be able to respect himself if he left her? If he stayed? Is he, at heart, a compassionate and dutiful son, or an adventurer and patriot?

Obviously, these are difficult questions. But the nature of their difficulty, and its relation to the question of the student's autonomy, depends on how the normative view of autonomy is understood. The view comes in subjective and objective versions. Both versions identify autonomy with a capacity for rationality. This is the essential claim of the normative view, full stop: that your practical reasoning and conduct is truly *up to you*—it is free, under your active control, a manifestation of your self-governance, something for which you are responsible in a certain privileged sense—just in case it is responsive to a certain class of reasons, defined relative to a certain class of values. (I'll call these, accordingly, the values with which you *identify*.) But they differ concerning the nature of these values. According to subjective versions—as defended, for instance, by Harry Frankfurt and Michael Bratman—the fundamental values with which you identify are subjective in the sense that their content is a function of facts about you that need not

obtain of other fully reasonable or rational agents.[1] Thus autonomy consists in responsiveness to *internal* reasons, in roughly the sense defended by Bernard Williams.[2] But according to objective versions—as defended, for instance, by Philip Pettit and Michael Smith, Joseph Raz, and (at one point) Susan Wolf—that is not so.[3] Rather, autonomy with the name requires the ability to respond to reasons *external* to your subjective values or commitments—or, as Wolf puts it the "ability cognitively and normatively to recognize and appreciate the world for what it is."[4]

On an extreme version of the objective view, the student's difficulty is purely epistemic. In asking himself what he should do, he is asking himself what the right thing to do is, all things considered, where this is independent of his individual personality and character. Similarly, in asking himself whether certain of his motives are worth acting on, or whether to value himself as one kind of person or another, he is trying to bring his judgments and self-conception into alignment with how they ought to be—how they would be, that is, if he were ideally rational, informed, and virtuous. The values with which he identifies, according to this view, are exhausted by the values that any ideally rational, informed, and virtuous person necessarily would have. His reasoning and conduct would be autonomous, therefore, only to the extent that they tracked these values in the right way—to the extent that they were suitably "reason-

1. For the most recent statements of Frankfurt's and Bratman's views, see *Taking Ourselves Seriously and Getting it Right* and *Structures of Agency*, respectively. Naturally, different subjective theorists differ the regarding the facts that fix the content of an agent's values. For Frankfurt they are facts about a subset of the agent's desires; for Bratman, they are facts about a subset of the agent's plans.

Note that the improvisational model I will eventually defend is meant to be as neutral as possible with respect to these and other subjective theories. It advances a thesis about how some of the agent's values get their content; if it is possible for Frankfurtian desires or Bratmanian plans to get their content in that way, the improvisational model is compatible with Frankfurt's and Bratman's views. (I see no reason to deny this possibility with respect to Bratmanian plans; I am less sure about desires. Certainly *Frankfurt* understood desires in a way that precluded it, which is why I do not find it unfair to introduce my view in opposition to his.)

2. See "Internal and External Reasons."

3. See e.g. "Freedom in Belief and Desire," *Engaging Reason*, and "Sanity and the Metaphysics of Responsibility," respectively. Note that so long as objective views allow that an agent's values may have irreducibly subjective *elements*, my view should be compatible with them, too.

4. "Sanity and the Metaphysics of Responsibility," p. 381.

responsive," as it were. (There is no need to broach the vexed question of what counts as "suitably" here.)

If it were indeterminate what an ideally rational, estimable, and virtuous person would do in the student's position, this may preclude him from acting autonomously. I do not know what it would be for this to be indeterminate, however. Perhaps it might be indeterminate if the student had *categorical* reason to do each thing—if his predicament were a dilemma in the strict, classical sense—in which case the conclusion that he could not act autonomously would be plausible. But I think the student's predicament is most interesting if it is imagined *not* to be a dilemma in this sense, and in this case, again, I find it hard to imagine what genuine rational indeterminacy would amount to. More plausibly, if the student were reasoning from the perspective of his ideally rational, estimable, and virtuous self, he would either recognize himself to have most reason, all things considered, to do one thing rather than another—in which case he would be acting autonomously if he did that thing—or he would recognize the reasons favoring each alternative to be roughly equal or on a par—in which case he would be acting autonomously if he did either. (Of course, the student's actual normative uncertainty may itself raise interesting questions about his autonomy, but these are irrelevant here.)

In any case, I do not think the extreme version of the objective view is very plausible, and I will assume—if only to explore what follows—that it is false. It is normal to care about what an ideally rational, estimable, and virtuous person necessarily would care about, but it is normal to care about other things, too. The values with which Sartre's student identifies may be "underdetermined," as Bratman observes, "by his intersubjectively accountable value judgments."[5] There is room for pluralism, for the recognition that some of one's personal values are at bottom just that. I may value certain projects or ideals as essential components of my

5. "Autonomy and Hierarchy," p. 166.

identity, without supposing it unreasonable that anyone relevantly similar to me should value different ones, or the same ones differently conceived. This need not require rejecting the objective view, of course—just its extreme version. The values with which an autonomous agent identifies may be constrained by what is objectively worth valuing without being exhausted by it.

Once it is accepted that the values with which we identify normally include irreducibly subjective elements, the possibility that the student's anguish may reflect a genuine indeterminacy in his reasons, rather than the mere obscurity of them, becomes more plausible and relevant. This is because our working conceptions of our values tend to be vague. Without the assumption that these conceptions aim to represent objective normative facts, it is dubious that there should always be a privileged specification of them, on which practical questions would have determinate answers.

It is easiest to see this by considering Sartre's student in more depth. As I am inclined to imagine him, he sincerely and sensibly loves his mother. We are told enough about her to make it understandable that he should not only feel protective of her (she is bereaved, lonely, and strained), but also admire her (it is implied that she was estranged from her husband partly because of his willingness to collaborate). He is also, on the other hand, a curious and humane young man in wartime Paris. He is probably very attached to ideals of forbearance, individualism, and adventure as well as ones of social justice and democracy, and thus powerfully drawn to the heroism of the Resistance on both moral and non-moral grounds. There is no reason to assume, however, that the content of his attitudes about all of these things—consciously represented or no—should have been any more structured or specific than was necessary for him to answer the concrete practical questions that confronted him.[6] Since, up to the point of his quandary, these questions may safely be supposed to have been by and large undemanding—

6. Cf. Henry Richardson, "Specifying Norms as a Way to Resolve Concrete Ethical Problems."

shall he buy *maman* the hat for her birthday, shall he go to the protest—one would expect their corresponding content to have been very unstructured and unspecific indeed.

Here it is useful to compare the probable character of Sartre's student's values with the way many philosophers of language conceive of indeterminacy in their analyses of vague predicates, like "bald."[7] "Bald," they argue, admits of multiple interpretations, or "sharpenings," each entailing a precise set of necessary and sufficient conditions for when a person is bald. When a claim is true on each sharpening—like "Jean-Luc Picard is bald"—we can say it's determinately true, or "supertrue." When a claim is false on each sharpening—like "James T. Kirk is bald"—we can say it's determinately false, or "superfalse." When it's true on some and false on others—like "Joe Biden is bald"—we can say its truth value is indeterminate.

Similarly, the student's values may be imagined to admit of multiple, mutually incompatible specifications, each constituting an interpretation of precisely what would be important to him in staying with his mother and going off to fight, and how the considerations in favor of each alternative are to be weighed. Sartre himself mentions some of these competing interpretations—the welfare of others may be important to him, but is it important to him in a sense that favors promoting the welfare of the concrete individual who depends on him, or the abstract but potentially much larger group he might aid by leaving her? To what extent is his inclination to stay with his mother to be endorsed as a manifestation of his real love for her, rather than rejected as mushy sentiment, or a shallowly conventional feeling of filial obligation?

Sartre claims these questions can't be answered—at least not in abstraction from what the student actually *does*. I am inclined to agree. Certainly, some specifications are better and more accurate than others—there may well be no admissible specification of the student's values on

7. Cf. e.g. Kit Fine, "Vagueness, Truth, and Logic."

which he has most reason to do anything *other* than one of the two alternatives in question.[8] And if the student's values were assumed to track some independent set of normative facts, it would be reasonable to think there would *always* be a uniquely best specification of them—the one that perfectly corresponded to those facts. But this, again, is just the assumption we are not making. Without it, it just seems unrealistic to me that the constraints on admissible specifications, whatever they are, should be robust enough to preclude the possibility, in the student's case, that there should be some admissible specifications of his values on which he has most (internal) reason, all things considered, to stay with his mother, some on which he has most reason to fight, and some (conceivably) on which his reasons for each are roughly equal on a par. If this is right, however, the student's reasons are indeterminate in the same way that the extensions of vague predicates are argued to be.

With the conspicuous exception of Harry Frankfurt, exponents of the normative view of autonomy have generally not considered the possibility and implications of such indeterminacy in depth. For his part, Frankfurt viewed it as *ambivalence*, which he understood as a kind of volitional fragmentation, a "disease of the will" incompatible with autonomous agency. "A person is ambivalent," he wrote, "only if he is indecisive concerning whether to be for or against a certain psychic position." That is:

> To the extent that a person is ambivalent, he really does not know what he wants. The ignorance or uncertainty differs from straightforwardly cognitive deficiency. There may be no information concerning his will that the ambivalent person lacks. He is volitionally inchoate and indeterminate.
>
> This is why ambivalence, like self-deception, is an enemy of truth. The ambivalent person does not hide from some truth or conceal it from himself; he does not prevent the truth from being known. Instead, his ambivalence stands in the way of there being a certain truth about him at all. He is inclined in one direction, and he is inclined in a contrary direction as well, and his attitude toward these inclinations is unsettled. Thus, it is true of him neither that he prefers one of the alternatives, nor that he prefers the other, nor that he likes them equally.[9]

8. For a plausible set of constraints on admissible specification, see Richardson, op. cit.

9. "The Faintest Passion," pp. 99-100.

It does not take much imagination to see how the ambivalence Frankfurt describes would be instantiated in the student's case. In asking himself what to do, the student asks, naturally enough, what is really important to him—how it is really important to him to live, and who it is really important to him to be. Also naturally enough, he finds himself completely unable to answer this question authoritatively. Again, this inability does not result from insufficient information. It also does not result (it's worth restating) from the kind of desire for some ultimate, independent vindication of one's personal values that Frankfurt and other subjective theorists reject as misguided. The student may aim at nothing more than being true to himself. (Objective theorists are free to stipulate that he already made sure that each alternative is morally permissible.) His problem is that there seems to be no fact of the matter as to what being true to himself would amount to. We can imagine the student coming up with different specifications of his values, given his initial jumble of inclinations and inchoate sense of self, and seeing no decisive grounds to favor any of them over the others. Each may be basically coherent, accurate to the data, and reasonably compelling. By what further standard is he to decide between them? I can think of no plausible answer.

Importantly, the absence of such a standard is not, in itself, a problem for Frankfurt's view. He is not committed to the claim that when you identify with a particular specification of an initially unstructured set of values, you are rationally committed to the specification you arrived at having been uniquely privileged.[10] In order to identify with one specification rather than another, it is enough for you to be wholeheartedly satisfied with it—that is, simply to desire to reason and act according to it, rather than any other candidate specification, and to have no

10. In fact it is not clear whether anything in Frankfurt's view implies that, when you identify with a more specific set of values than you did at an earlier time, you are rationally committed to the later set of values even counting as an *admissible* specification of the earlier set (such that the rational demands implied by the later one are consistent with those implied by the former, and so on), as opposed to a completely different set of values entirely. But since more sophisticated subjectivist theories (such as Bratman's) do not share this problem, I will ignore it.

desires to the contrary. In order for Sartre's student to be capable of acting autonomously, then, he must, on Frankfurt's view, come to have such a desire. His ambivalence issues from his not having it. But note the psychological change the student would undergo in coming to have it would not be a rational change. By this, I do not mean that the student would just have to arbitrarily plump among the admissible specifications in deliberation, but rather that his coming to identify with one specification rather than another, or none at all, is not, for Frankfurt, distinctively a matter of deliberation and choice at all. "A person cannot make himself volitionally determinate," he writes, "and thereby create a truth where there was none before, merely by an act of will."[11] To be sure, it is possible that, upon entertaining a given candidate specification in deliberation, the student might become inexorably drawn to it, and inexorably satisfied with being so drawn. But the fact that this change should happen to occur in deliberation would be pure accident. It would not be under his deliberative or voluntary control, and it could have had all sorts of other causes.

For Frankfurt, then, no less than for objective theorists, autonomous agency is a matter of discovery, rather than invention. Our individual values are, for him, no more up to us to choose for ourselves than is the universal normative reality to which objective theorists advert. We take active control over our agency by finding out what the values with which we identify are and acceding to their demands, and if there is no prior truth about what those values are, then we cannot act autonomously. Needless to say, Sartre viewed the matter differently. It is indeed the case, he held, that our choices are not rationally determined in advance, either subjectively or objectively. Yet it would be in bad faith if we took the absence of such determination to excuse us from taking full responsibility for our actions, as in some deep and distinctive sense our own. We must *invent* the values with which we identify, not discover them.

11. "The Faintest Passion," p. 100.

Though I hardly want to invoke Sartre's full-blown metaphysics of agency and value, I am inclined to agree with him here, at least with regard to his student's case. The indeterminacy I have attributed to the student is probably very common, and similarly exhibited in many of our most important and formative choices. To be fair, Frankfurt does not dispute its pervasiveness: he is clear that ambivalence, to some degree or other, is something we all have to live with. But I think his view of its consequences for our agency is too extreme. When we make choices under conditions of indeterminacy, we generally feel unsettled and unsure of ourselves—and sometimes even anguished—if we make them reflectively or self-consciously. (It is probably worth keeping in mind that we often do *not* make them reflectively or self-consciously, however.) In many cases we feel it to be impossible to fully explain our reasons for making them. But for all that we tend to regard ourselves as more than just bumbling about. Certainly we *see* our choices as things for which we can and should take an important kind of responsibility, closely linked to the significance they have in our lives. I do not want to try to say much about this kind of responsibility until I have my alternative view in place, but to get the ball rolling, we might look more closely at the student's anguish.

In *Being and Nothingness*, Sartre writes that in committing to a temporally extended project, "I await myself in the future, … I 'make an appointment with myself on the other side of that hour, of that day, or of that month.' Anguish is the fear of not finding myself at that appointment, of no longer even wishing to bring myself there."[12] In the vocabulary of Anglophone practical philosophy, the student sees himself as *answerable* to his future self, normatively speaking: he's concerned about the prospect of his future regret. Crucially, however, he isn't—or shouldn't be—concerned about this *unconditionally*. He might well come to regret his choice if he undergoes

12. *Being and Nothingness*, p. 73. Note the difference between these remarks and Sartre's discussion of anguish in "Existentialism is a Humanism," wherein it is conceived, much less plausibly, as a sense of answerability to all of humanity.

a fundamental change in values. But while this would be unfortunate, its possibility should not check his resolve. In Jonathan Franzen's novel *Freedom*, a character named Lalitha describes her intention to have her Fallopian tubes tied, out of both her political beliefs regarding overpopulation and her personal desire to live without children. When it is observed that her future self might not share these attitudes, Lalitha replies, surely rightly: "Then *fuck* my future self. If it wants to reproduce, I already have no respect for it."[13] (Much the same can be said on behalf of the young nobleman Derek Parfit describes, to take the example better known to philosophers.) This implies that Sartre's student holds himself answerable not for adhering to *whatever* values he might eventually come to have, but only for adhering to values relevantly similar to his present ones. But how could he sensibly do this in the course of making his choice, unless, despite his present condition of rational indeterminacy, he could nevertheless regard that choice as putatively embodying his values?

This should not be confused with an argument for voluntarism. Voluntarists claim that we can indeed invent our values: when our existing reasons fail to determine what we ought to do, we can create new ones through rationally undetermined, self-justifying acts of will.[14] If it is indeterminate whether it is most important to Sartre's student to stay with his mother or to fight, he can simply will himself to be someone to whom one is more important than the other, and *ipso facto* become such a person. But self-creation cannot be so easy. In claiming that Sartre's student can determine his reasons *just* by willing—at the moment of his choice, as if stipulating—that

13. *Freedom*, p. 308. Note that Lalitha's case differs from Parfit's nobleman's in that the former does not support the view—which some have defended with respect to the latter—that it is appropriate or sensible to disregard the values of a future self only if those values are objectively defective (or if you justifiably believe them to be so). (See e.g. Christine Korsgaard's treatment of the case, in *Self-Constitution*.) Lalitha may quite sensibly reject such attitudes in *her* future self without supposing *other* people who have them to be equally unworthy of respect.

14. See e.g. Robert Nozick, *Philosophical Explanations*, and Ruth Chang, "Voluntarist Reasons and the Sources of Normativity."

they be thus and so, the voluntarist is, if anything, even less able to account for his anguished sense of responsibility than Frankfurt is. As Frankfurt himself explains:

> The concept of reality is fundamentally the concept of something which is independent of our wishes and by which we are therefore constrained. Thus, reality cannot be under our absolute and unmediated voluntary control. The existence and character of what is real are necessarily indifferent to mere acts of our will.
>
> Now this must hold for the reality of the will itself. A person's will is real only if its character is not absolutely up to him. It cannot be unconditionally within his power to determine what his will is to be, as it is within the unconstrained power of an author of fiction to render determinate—in whatever way he likes—the volitional characteristics of the people in his stories.[15]

If Sartre's student takes himself to have most reason to fight, it is essential to his choice—as it is to all species of norm-governed activity—that it admit of the conceptual possibility of error. This is not a merely technical point. The specter of the student's regret makes it vivid. It implies the belief that his present choice might turn out to be mistaken, that it might fail to conform to his values. This belief would be unwarranted if his choice could not thus fail. And it *necessarily* could not thus fail if, as voluntarism implies, the student's choice conformed to his values only because his *values* automatically conformed to his *choice*.

* * *

At the end of the day, I think Frankfurt and the voluntarists have the same problem—a problem shared by every major version of the normative view of autonomy I know of.[16] They implicitly view the values with which we identify, and to which we hold ourselves answerable for adhering, as fixed, static things. This isn't to say they think, crazily, that you can't change the values with which you identify while remaining the same person. But they view the *content* of those values, at a given time, as exclusively a function of facts that obtain of you *at* that time, or prior to it. To know all there is to know about your values, that is, all we would need is a snapshot of the

15. "The Faintest Passion," pp. 100-101.

16. Or, more exactly, no major version of the normative view I know of actually *rejects* the view I claim as problematic. (To be fair, not all explicitly accept it, or are even committed to doing so.)

relevant facts—facts about your attitudes and their relations to one another, for Frankfurt; facts about your acts of will, for the voluntarists. I now want to argue that we can understand Sartre's student's sense of responsibility to himself—and why that sense of responsibility is something normative views of autonomy should countenance—if we took a different view of his values. On this view, Sartre's student identifies with, and sees his choice as to be justified in terms of, a set of values whose content is still in the process of being determined, and so depends on future facts as much as present and past ones. To make this possibility understandable, it will help to draw an analogy to musical improvisation.

3 Improvisation as a model of self-creation

Improvisation differs from other species of rational agency in that improvisers refine the ends they pursue—that is, the norms they are committed to—as they go. When you improvise, you act in the ways you take to be appropriate without necessarily being able to explain why. Yet you don't see your actions as random, but rather as parts of a process of working out an expressive musical performance, say—or an overall way of life—that explains why each of the actions that constitutes it is appropriate in relation to the whole. To see how this might go in the music case, consider how Keith Richards recalls the improvisation that went into *Exile on Main Street*:

> We were prolific. We felt then that it was impossible that we couldn't come up with something every day or every two days. That was what we did, and even if it was the bare bones of a riff, it was something to go on, and then while they were trying to get the sound on it or we were trying to shape the riff, the song would fall into place of its own volition. Once you're on a roll with the first few chords, the first idea of the rhythm, you can figure out other things, like does it need a bridge in the middle, later. It was living on a knife edge as far as that's concerned. There was no preparation. But that's not the point; that's rock and roll. The idea was to make the bare bones of a riff, snap the drums in, and see what happens. And it was the immediacy of it that in retrospect made it even more interesting. There was no time for much reflection, for plowing the field twice. It was "It goes like this" and see what comes out. And this is when you realize that with a good band, you only really need a little sparkle of an idea, and before the evening's over it will be a beautiful thing.[17]

17. *Life*, p. 306.

Think of a spontaneous decision to play a riff because it strikes you, right then, as especially *expressive*. Decisions like this seem to occupy an intermediate position between rationally determined conclusions of deliberation and one-off random acts. On the one hand, you make them because they feel like the way to go at the time, not because anything in your performance demands them. Relative to everything you've played up to that point, any number of other riffs may have been equally musically good, or better; you might no less reasonably have played any of them. (The qualification "up to that point" is critical here, as we'll see in a moment.) On the other hand, you make them because they *feel like the way to go* at the time—the particular riff you pick feels right in some way, as more than just a matter of indifference.

We can plausibly capture this middle ground by recognizing how a riff can strike you as especially expressive of *something* without your being able to say what it's expressive *of*. It's not necessary for you to be in the possession, or suppose yourself to be in the possession, of any facts that would explain why that specific riff was the right thing to do, given your end. The *causal* explanation of your decision to play it, as opposed to anything else you might equally well have played, would cite all sorts of totally incidental features of your psychology and circumstances. (Imagining a causal explanation of *Keith Richards's* decisions, the mind boggles.) Yet the whole point of improvisation is that in such a case you're *not* treating your decision as incidental. Parts of a musical performance aren't expressive in isolation. The feel of a riff is colored by the past playing that anticipates it and the future playing that integrates and elaborates on it; on its own, it would fall flat. When you take your riff to be expressive, you're implicitly situating it, normatively speaking, into a larger context. You see it as part of a process of expressing a specific musical idea. As you work out the contours of that idea, you're simultaneously working out what the riff you took to be *expressive* actually helps *express*.

This lets us give substance to the idea of making up your ends as you go along. It's plausible—as a general thesis about rational agency—that when you act for a reason, rather than arbitrarily, you necessarily presuppose that there is an explanation why the action you perform is appropriate.[18] What makes improvisation different from other species of norm-governed activity is that when you improvise, the requisite justifying explanation essentially depends on the particular judgments and actions you perform over the course of the improvisation.

How can a justifying explanation for a particular response essentially depend on a set of others? The question arises, I think, because it can be tempting to assume that to give a justifying explanation for a response is to subsume it under some general rule or principle: I can explain why it is appropriate to say "10" after "8" in a simple counting game by citing a rule of adding two. But it is a mistake to think that *all* justifying explanations must be like this.[19] Some perfectly good ones simply cite other cases, thereby making the appropriateness of the response at hand intelligible as part of a natural pattern. In the counting game, for instance, we can say: "Look, we know '6' called for '8,' '4' for '6,' and '2' for '4.' So it's only natural that '8' should call for '10.'" In this context, of course, it's obvious what the general rule in force would be, but in others—including, prominently, aesthetic ones—it may not be. When you explain why a painter's decision to represent a face in a certain way helped communicate a vivid sense of the subject's personality, it's not always necessary to give an explicit interpretation of the personality the subject was represented to have. Sometimes it's even impossible. ("If you could say it in words," Hopper said, "there would be no reason to paint.") But you can still provide an illuminating

18. For an influential presentation of the concept of an "explanation why" in the sense I have in mind, see John Broome, "Reasons.". There, Broome argues that this concept is *prior* to the concept of a normative reason, but it is obviously unnecessary to accept Broome's conclusion in order to admit a necessary relation between the two.

19. Arguably, at least some justifying explanations *cannot* be like this, in light of the puzzle about rule-following Kripke famously attributed to Wittgenstein in *Wittgenstein on Rules and Private Language*. Cf. John McDowell, "Non-cognitivism and Rule-Following," and Hannah Ginsborg, "Primitive Normativity and Skepticism About Rules," the latter of whom likewise takes inspiration from aesthetics.

explanation of the painter's decision by referring to other parts of the portrait, in the context of which the decision in question just *worked*.

So when you take a certain riff to be expressive, you likewise presuppose that your decision will admit of a justifying explanation in terms of its significance to your performance as a whole, as the cogent expression of a unified musical idea. But the key point is that you needn't have that explanation *at the time*, because the nature of the riff's significance has *yet to be determined*. It depends on how what you're playing now coheres with what you'll play in the future. There are lots of ways in which you might go on to elaborate on the riff and integrate it with everything else. What it ends up expressing depends on the musical idea it ends up being a partial expression of, and which in turn depends on the direction you end up take your improvisation in. It's even possible that if you go on to ignore the riff completely, or end up with a total mess, it won't turn out to express anything at all. In such a case, the presupposition you made in deciding on the riff—that you'd end up with an explanation of its having been appropriate—will be falsified, and your decision to play it will therefore turn out to be a mistake.[20]

Thus when you improvise, you pursue an end whose content depends, epistemically and ontologically, on the actions you actually take over the course of pursuing it: relative to your end, you have reason to do something just to the extent that it admits of justifying explanation in terms of the other actions you perform that are likewise explicable themselves. As such the vague feeling that a certain riff is somehow the way to go at the time doesn't register your conformity to an existing obligation, but rather your acceptance of a commitment. You could, after all, have played something else, and thereby gone on to express a musical idea of which it—rather than what you actually played—was a necessary constituent. Yet precisely because you regard yourself

20. So randomly plunking at piano keys isn't improvising, because barring some monkey-writing-Shakespeare sort of accident, you could only produce a disorderly jumble of notes. There will be no explanation why any of them shouldn't have been different, and you won't have expressed anything.

as aiming to express a specific musical idea—though one still in the process of being shaped—you don't regard the decision you've *actually* made as rationally on a par with the one you *could* have made but didn't. You rather treat it as justified relative to the end you're thereby determining for yourself, and hence to be taken into account in your ongoing practical reasoning as bearing on what you have reason to do in other cases (in virtue of its explanatory interdependence with it). So it makes perfect sense for you to take yourself to have reason to play the specific riff you do, as partly constitutive of expressing your musical idea. It's just that because your musical idea is essentially something you're in the process of working out, so are your reasons.

4 Deep improvisation, or, Sartre's student and Williams's Gauguin

It's now time to return to Sartre's student. Let's review. In §2, I used the case of the student to raise a puzzle for normative views of autonomy. These views claim that to take responsibility for a choice as autonomous is to take yourself to have sufficient reason to make it, given the values you identify with. This may seem to imply, as Frankfurt supposed, that autonomy is impossible under conditions of rational indeterminacy. It is important, I claimed, for agents like Sartre's student to take themselves to be choosing autonomously under such conditions and in full recognition of them, and to be potentially correct in doing so. The question was how this was possible. Voluntarists answer this question by arguing that taking yourself to have most reason to do something, under conditions of prior indeterminacy, makes it so. While it should now be clear that voluntarism has an element of truth to it, I claimed, following Frankfurt, that it makes self-creation too easy. Normative standards are meaningful as such only insofar as it is possible to get them wrong.

Conceiving the student's values on the model of musical ideas being improvised gives us a better answer.[21] (For the sake of a name, call this *deep* improvisation.)[22] Rational indeterminacy prevents you from being warranted in taking yourself to have sufficient reason to do something only if taking yourself to have sufficient reason to do something necessarily involves making a *judgment* about the state of those reasons as they stand—namely, the judgment that it is (now) true that you have sufficient reason to do that thing. (For since your reasons are indeterminate, it cannot be thus true.) But the possibility of improvising shows that taking yourself to have sufficient reason to do something does *not* necessarily involve this. If your reasons are ones you are making up as you go, it rather involves accepting a commitment to treating what you do as to be taken into account in your ongoing practical reason and conduct, as something that is to help explain, and be explained by, what you likewise have reason to do on other occasions. Insofar as your present choice actually turns out to stand in this mutual explanatory relation with your past and future choices, you turn out to be warranted in taking yourself to have sufficient had reason to make it.[23] But it might not stand in this mutual explanatory relation. Whether it does depends on your own efforts in taking your choices over time seriously, as partially embodying the values you are now trying to live by, and on your luck in having the opportunity to choose in ways that allow this. Frankfurt writes, against voluntarism, that we "are not fictitious characters, who have

21. Importantly, I'm not arguing that the contents of *all* the values with which a person can identify are necessarily up to that person to improvise. In fact, it is plausible that the contents of at least some such values (such as specifically moral values) necessarily are not.

22. This terminology registers a debt to Charles Taylor, who has used the phrase "deep reflection" to refer to a process through which one simultaneously articulates and shapes the values with which one identifies by critically interpreting one's evaluations in light of one's "deepest unstructured sense of what is important." ("Responsibility for Self," p. 41)

23. This doesn't mean deep improvisers are forever bound to their pasts. The requirement that improvisers act in ways that mutually explain each other only holds among actions they *in fact* have reason to perform. So you may sometimes be justified in rejecting some of your past (or even predictable future) actions, and hence in ceasing to treat them as explanatory and to be explained. But nor are deep improvisers free to be entirely capricious: in rejecting some subset of your actions, you commit to having an explanation (from the perspective you thereby come to inhabit) why the rejection itself was called for, and such an explanation may be difficult to come by.

sovereign authors; nor are we gods, who can be authors of more than fiction. Therefore we cannot be authors of ourselves."[24] If I am right, Frankfurt's premises are true but his conclusion is false. We *can* be authors of ourselves, but not sovereign ones. Our ability to create ourselves is constrained by the real and demanding requirement that the selves we create be *recognizable as* particular selves persisting over time, where this means being embodied in coherent patterns of agency.

Now, the idea of deep improvisation might be resisted on two basic grounds. First, someone might reject the premise the whole puzzle began with: that rational indeterminacy should be compatible with autonomy, as conceived by the normative view. What would be lost—what significant feature of our agency would be neglected—if all we could do was throw up our hands, deal with the consequences of our choices as best we could, and hope to become less ambivalent later on? Second, even assuming an answer to this question, it might be asked whether the substance of my view is motivated. I have claimed that what makes improvisation intelligible as a form of norm-governed activity is the formal requirement that choices over time cohere. This requirement is plausible enough in the aesthetic case, since musical performances can't express anything if they don't cohere. But most of us don't value *everything we do* as a form of self-expression. So why think the coherence requirement generalizes?

I will tackle the last question first. My answer is that while you certainly needn't care about self-*expression*, you can't meaningfully identify with values that couldn't support a kind of self-*recognition*, as a particular self persisting over time. Now, I'm not saying that you have to value self-recognition more than anything else, or even that it can't be perfectly rational to identify with values that make it *very unlikely*, in point of empirical fact, that you actually *will* be able to recognize yourself. I'm just saying that, as a condition on the possibility of identification, it must

24. "The Faintest Passion," p. 101.

be the case that if (perhaps *per impossible*) you *did* manage to reason and act in all the ways your values required of you, your conduct *would* be recognizable as embodying a unified set of values. And any values that meet this condition, at least if they can support autonomy under conditions of rational indeterminacy, must meet the coherence requirement.

Here's why. We've seen that in order to choose autonomously under conditions of rational indeterminacy, you need to identify with values you're making up as you go—ones whose content is epistemically and ontologically prior to the choices they govern. But in order to meaningfully identify with a value, you need to have some principled means of determining what it requires in particular situations. Recall Lalitha's case, from *Freedom*. It is, again, central to her integrity that she be able to reject, as a kind of self-betrayal, any possible future desire to have children. Any plausible conception of identification must be able to countenance this. But on what grounds could she reject this desire, as opposed to any other one? I can only see two possibilities. First, if the content of her values were accessible to her prior to her choices in particular cases, she could deduce what her values required from it. But if she is still working out what her values are, she cannot *per hypothesis* have such access. This leaves the second possibility: she could infer what her values require in the present case from what they require in others. But *that* possibility requires that her values meet the coherence requirement. For how could her present history enable her to reject her possible future desire, if not its clear lack of fit with any coherent trajectory by which her values might evolve? More generally, recall that all I mean by coherence here is that a set of choices be intelligible as falling under a unified norm, such that the appropriateness of each relative to the norm is naturally inferable from that of the others. And in this light, the idea of identifying with values that *don't* meet the coherence requirement should seem pretty bizarre. It would amount to identifying with values that call for you to choose in ways that aren't intelligible as embodying *any* particular set of values at all. What could possibly

be the point? Such values would leave people like Lalitha with no ways of thinking of themselves as having particular characters, identified with some patterns of desire and not others, that were not unaccountably ad hoc. They would not support any sense of thematic unity or larger meaning in one's activities, any means of making oneself of oneself as a particular sort of person persisting over time.[25] Critical reflection would be otiose, and identification would be an empty concept.

The first question is harder. The best answer I can think of appeals to the value of throwing yourself into a project—of investing your whole self in it, indeterminacy and contingency be damned. This may be much of what Sartre had in mind in upholding authenticity as an existentialist ideal. It is also a value conspicuous in art, especially jazz.[26] But its best illustration, at least in the present context, may be in Bernard Williams's fictionalized Gauguin.

Williams introduced Gauguin in order "to explore and uphold the claim that in such a situation [viz., his abandoning his family in order to pursue his art] the only thing that will justify his choice will be success itself." That is:

> The justification, if there is to be one, will be essentially retrospective. Gauguin could not do something which is thought to be essential to rationality and to the notion of justification itself, which is that one should be in a position to apply the justifying consideration at the time of the choice and in advance of knowing whether one was right (in the sense of coming out right).[27]

One reason why Williams's thesis has, for all its notoriety, found little acceptance among its readers may be that Williams is frustratingly opaque about the kind of justification he has in

---

25. Cf. David Wiggins, "Truth, Invention, and the Meaning of Life," in his *Needs, Values, Truth: Essays in the Philosophy of Value* (third edition, Oxford, 2003), in which the mutual explanatory support I identify with coherence is upheld as characteristic of a meaningful life.

26. This seems not to have been lost on Sartre. At the end of *Nausea*, the protagonist realizes that a certain kind of serious, authentic life is possible even in the absence of authoritative normative standards only upon listening to a blues record.

27. "Moral Luck," pp. 23 and 24, respectively.

mind. He wisely refrains from describing it as a species of moral justification when he introduces it, but when he returns to the case at the end of the essay, he argues (against Thomas Nagel) that if the justification in question is not strictly speaking moral, it is at least closely akin to it. But that is unconvincing. Presumably it is a rotten thing to abandon people who justifiably depend on you. If you manage to realize your dreams by doing it, this might make it good for you, but it does not make it any less rotten. (If realizing your dreams means producing great art, it might also make it good for the world, but this cannot be germane to the justification at issue. If nobody, or only a few people, ever saw Gauguin's paintings, this would not make his choice unjustified in the manner his failure would.) But if we viewed the question of Gauguin's justification as instead turning on the personal values with which he identified, Williams's claim starts to look a lot more plausible.

That Williams's Gauguin should exhibit to the same kind of rational indeterminacy as Sartre's student may come as a surprise. Unlike the student, Gauguin seems to know what he wants. He seems like a paragon of Frankfurtian wholeheartedness. But this reading is superficial. If Gauguin does not feel unsure of himself, he should. For consider what his relations to art on the one hand, and his family on the other, must have been like at the time of his choice. He was not, as Williams imagines him, in any sense entitled to then regard himself as a great artist. This does not only bear on his evidence that his project will succeed, but on the significance he could reasonably attribute to it. However intensely he might want to paint, he may reasonably believe that desires like his are usually not to be acted on. Many other men relevantly similar to him may well have wanted to remark on similarly romantic endeavors no less intensely, and most of them—we may allow Gauguin to think—turned out to be pathetic failures who should have known better. In this spirit, Williams considers and rejects the possibility that Gauguin should decide what to do by reference to rules:

> [The rule that 'one is justified in deciding to neglect other claims for art…'] '…if one is convinced that one is a great creative artist' will serve to make obstinacy and self-delusion conditions of justification, while 'if one is reasonably convinced that one is a great creative artist' is, if anything, worse. What is reasonable conviction supposed to be in such a case? Should Gauguin consult professors of art? The absurdity of such riders surely expresses the absurdity in the whole enterprise of trying to find a place for such cases within the rules.[28]

Williams is talking about *moral* rules here, but it seems to me that his point applies to prudential rules, too, and more broadly to any general principles or policies by which a sane person might deliberate. I can think of no such principle, moral or otherwise, on which the intensity and insistence of Gauguin's desire to give up everything for painting could count as a sufficient reason to do it, given the gravity of the risks associated with the project and the paucity of the evidence that it will succeed. Consider, for instance, the following candidate. On one reading of Frankfurt's view of practical rationality—although perhaps not an entirely fair one—Gauguin would be justified in acting on his desire to paint if he wanted to be moved by it more than he wanted to be moved by any other desire, and he couldn't help feeling this way. But this is no more plausible as a source of antecedently applicable justifying conditions than any of the rules Williams considers. If Gauguin's project turns out to be a failure, wouldn't it be understandable for him instead to think he should have found some way to eliminate or weaken this desire (by going to a psychologist, say)?

If this is right, it follows that if Gauguin is indeed going to take himself to have sufficient reason to paint, it can only be by taking a leap of faith. And I cannot imagine him adopting this project, or being motivated to pursue it to the fullest, *without* taking himself to have sufficient reason to paint. One reason why Williams's example works so well is that the historical Gauguin's biography and body of work, taken together, really do constitute a strikingly unified picture of a complicated and provocative worldview manifested in a career of artistic genius. But it's doubtful that he could have developed this worldview except by improvising. Yet it's neither

28. "Moral Luck," p. 24.

plausible nor relevant to suppose that Gauguin had anything like the specific ideals he turned out to embody in mind, however subconsciously or implicitly, when he first decided to throw himself into his project. There's no reason why, for instance, had Gauguin not gone to Tahiti, he would necessarily have become quite the sort of person who would have opposed (some of) the injustices of French colonialism to the extent that he did, or would have developed his interest in primitivism and mysticism into a consuming preoccupation with certain fundamental questions of human existence—even though these turned out to be essential to the importance a life of art turned out to have for him. He might have gone to Siberia, or stayed in Paris, and became a artist with a different sensibility, albeit perhaps an equally profound and unified one. But—crucially—it's doubtful that he could have developed a profound and unified artistic sensibility in *any* of these situations if he was not propelled forward by the inchoate, unstructured, not-determinately-justified conviction that a life of art really was—somehow—profoundly important to him, and if he didn't interpret and develop his responses to whatever circumstances he was presented with through its lens. What he has to ask himself, at each stage of the process, is how the various activities he finds himself engaged in, in light of how he's taken it to be important to him to act so far, can be fit together into a life so richly and fully animated by the significance of painting to him as to make his initial, momentous decision understandable. Since such a life could only be that of a great creative artist, Williams's claim that Gauguin's choice could only be justified by its success turns out to be right.

5 Conclusion

Let's review. I began with the question of autonomous agency under conditions of rational indeterminacy. Normative views say that autonomous agency is *rational* agency: acting autonomously requires doing what you have, or take yourself to have, sufficient reason to do, all

things considered, given the values with which you ultimately identify. It may therefore seem as though autonomous agency is impossible if it's indeterminate what you have sufficient reason to do, and this is exactly the conclusion that Frankfurt drew. But this conclusion is false. Rational indeterminacy doesn't mean you can't act autonomously. It may mean you just have to *improvise*—which means taking a leap of faith, making up your reasons you go, and trusting that you *will* be able to make sense of your action as embodying the identity you're creating for yourself.

But what turns on this? Why is improvising something we find ourselves *having* to do? Why does the possibility of it *matter*? In the following chapters, I'll offer the beginning of an answer.

2

Loving someone in particular

1 A somewhat dubious case study

Let me begin with a scene from one of the most famous—if problematic—novels about love ever written. In *Wuthering Heights*, Catherine Earnshaw consents to marry Edgar Linton, a perfectly eligible match. But she is ambivalent about it. So she asks Ellen Dean, her longtime servant and confidante, whether she ought to have done so. The following conversation (related from Ellen's perspective) ensues.

> "There are many things to be considered before that question can be answered properly," I said sententiously. "First and foremost, do you love Mr. Edgar?"
>
> "Who can help it? Of course I do," she answered.
>
> Then I put her through the following catechism: for a girl of twenty-two, it was not injudicious.
>
> "Why do you love him, Miss Cathy?"
>
> "Nonsense, I do—that's sufficient."
>
> "By no means; you must say why."
>
> "Well, because he is handsome, and pleasant to be with."
>
> "Bad!" was my commentary.
>
> "And because he is young and cheerful."
>
> "Bad, still."
>
> "And because he loves me."
>
> "Indifferent, coming there."
>
> "And he will be rich, and I shall like to be the greatest woman of the neighborhood, and I shall be proud of having such a husband."
>
> "Worst of all! And now, say how you love him."
>
> "As everybody loves—You're silly, Nelly."
>
> "Not at all—Answer."
>
> "I love the ground under his feet, and the air over his head, and everything he touches, and every word he says—I love all his looks, and all his actions, and him entirely, and altogether. There now!"
>
> "And why?"
>
> "Nay—you are making a jest of it; it is exceedingly ill-natured! It's no jest to me!" said the young lady, scowling, and turning her face to the fire.
>
> "I'm very far from jesting, Miss Catherine," I replied. "You love Mr. Edgar, because he is handsome, and young, and cheerful, and rich, and loves you. The last, however, goes for nothing: you would love him without that, probably, and with it you wouldn't, unless he possessed the four former attractions."

"No, to be sure not—I should only pity him—hate him, perhaps, if he were ugly, and a clown."

"But there are several other handsome, rich young men in the world; handsomer, possibly, and richer than he is. What should hinder you from loving them?"

"If there be any, they are out of my way—I've seen none like Edgar."

"You may see some; and he won't always be handsome, and young, and may not always be rich."

"He is now; and I have only to do with the present—I wish you would speak rationally."

"Well, that settles it—if you have only to do with the present, marry Mr. Linton."

Ellen's "catechism" strikingly anticipates the issues on which contemporary philosophical discussions of love focus, and the features that leading accounts defend as necessary conditions for loving someone as a particular individual.[1] Harry Frankfurt, for instance, insists that someone loved in the best sense is valued as *irreplaceable*: if Catherine really loved Linton, it would not be a matter of indifference to her that she love him in particular, as opposed to anyone else with the same attractions. J. David Velleman stresses that love should involve a special *openness* to beloveds as they are in themselves, not just insofar as they serve your independent purposes or meet some prior standard. Catherine shouldn't love Linton just because he pleased her, or satisfied her vanity, and it's impossible to see her claim to love him indiscriminately and in total as other than a sarcastic parody of really loving attention. And Niko Kolodny emphasizes that love should be *constant*: it should endure through a very wide range of possible changes in a beloved. It shouldn't lapse, as Catherine's would, when beloveds lose their looks, youth, cheer, or wealth.[2]

Even more strikingly, however, the ideal of love the novel presents in opposition to the defective view represented in Catherine's initial responses is one that none of these philosophers can explain or even accommodate. For the real source of Catherine's ambivalence is that, as she

1. As the word is used in contemporary English, many things other than persons can be loved: animals, inanimate objects, institutions, activities, abstract ideas, deities, and so on. Though my discussion touches on love for some of these things at points, I assume as a working hypothesis that there is a distinct, philosophically interesting species of love essentially focused on particular persons. It is with this species of love that the following is concerned.

2. Frankfurt gives his theory its signature statement in "On Caring," and its most refined one in *Taking Ourselves Seriously and Getting It Right*. Velleman presents his view in "Love as a Moral Emotion" and elaborates it in "Beyond Price." Kolodny's proposal is in "Love as Valuing a Relationship."

well knows, she doesn't really love Linton at all. She *really* loves Heathcliff, the darkly romantic foundling. And she loves him for a very different kind of reason:

> "...not because he's handsome, Nelly, but because he's more myself than I am. Whatever our souls are made of, his and mine are the same, and Linton's is as different as a moonbeam from lightning, or frost from fire."

Catherine's answer raises a puzzle. Why should qualities of Heathcliff's soul—or, less metaphorically, of his identity or character—do any better by the standards of Ellen's catechism than any of the qualities Catherine cited in Linton's case? Aren't the values with which Heathcliff identifies just as repeatable, in principle, as Linton's handsomeness or wealth, and just as liable to undergo changes that real love should survive? Isn't their significance to Catherine just as circumscribed by her private interests and criteria, if not more so? Perhaps impressed by such questions, Frankfurt, Velleman, and Kolodny all defend theories on which the qualities of one's character and values are indeed no more suited to serve as reasons for love than any other quality of one's person. In this, they represent a broad consensus among analytic philosophers on love. But I will argue below that these philosophers are wrong, and Catherine is right.

My defense of Catherine's kind of love will proceed in two stages. In the first, I argue, against Frankfurt, Velleman, and Kolodny, for the possibility and importance of a species of interpersonal love evaluatively grounded in attractive qualities of the beloved. Thus, Frankfurt denies that love is a rational response to value to begin with, Velleman argues that it is a rationally optional response to a value that all persons (by definition) share equally, and Kolodny argues that it is a response to the value of the relationship you have to your beloved. But I argue, first, that there must be reasons for love; second, that these reasons must (at least in some cases) be selective; and, finally, that these reasons must ultimately derive (again, at least in some cases) not from the types of relationships you have to beloveds, but from what beloveds themselves are like. These theories, then, leave a void that the ideal of Catherine's love promises to fill.

Still, I think each theory gets something important right. Taken together, they show that the irreplaceability, openness, and constancy on which Ellen implicitly insists really are necessary to the best kind of love. Thus, Catherine's answer can fulfill its promise only if the puzzle it raises can be solved. In the second half of the chapter, I argue that it can. It's possible, and plausible, to conceive of one's identity as an agent as having special structural features that enable it, distinctively, to support a form of love that fully satisfies Ellen's catechism. In a way, this turn to basic features of agency and valuing should be unsurprising: the philosophical questions raised by the phenomenon of loving someone as an individual turn out to be questions about the nature of individuality itself.

In the previous chapter, I argued that we work out our identities by improvising, the way musicians work out the ideas they're trying to express. Lovers improvise *together*, as partners: when Catherine says she loves Heathcliff for a soul like hers, she's saying the same kind of thing that musicians might, when they take one another to want to express the same things. More specifically, I propose, to love someone in particular is to view that person as creating an identity that is somehow importantly like your own, in a way that makes your beloved someone appropriate for you to create yourself together with. But because your reasons for love are grounded in features of your and your beloved's identities that are in the process of being determined, those reasons persist throughout that process, and call for essentially open-ended forms of interested attention and emotional vulnerability. Further, they make you and your beloved irreplaceably valuable to one another, since someone you are creating your values together with can share those values in a way that nobody else can.

2 Loving someone for no reason

Perhaps the most glaring problem with Catherine's reasons for loving Linton is that they make him too easy to replace. What qualifies him as a suitable beloved is simply that he is a member of the general class of handsome, cheerful, rich young men. Anyone other member of that class would have done just as well. "With regard to what we love," however, Harry Frankfurt observes that "that sort of indifference to the identity of the object of concern is out of the question. Substituting some other object for the beloved is not an acceptable and perhaps not even an intelligible option. The significance to the lover of what he loves is not that of an exemplar; its importance to him is not generic, but ineluctably particular."[3] If you love someone as a particular individual, then, you value your beloved as *irreplaceable*.

This means, at a minimum, that it must be important to you that you love the particular person you do. Now, the simplest way to account for this importance would be to hold that reasons for love are perfectly particular themselves. Following a direct (if flat-footed) interpretation of Montaigne's famous non-explanation of his love—"because it was he, because it was I"—Catherine would thus have reason to love Heathcliff but not Linton, but that reason would be primitive and hence inexplicable. But this is a non-starter. "The beloved's bare identity," as Kolodny neatly explains, "cannot serve as a reason for loving her. To say 'She is Jane' is simply to identify a particular with itself. It is to say nothing about that particular that might explain why a specific response to it is called for."[4] We might as well say love has no reasons at all.

Such is Frankfurt's view. Love, he argues, "is a particular mode of caring. It is an involuntary, nonutilitarian, rigidly focused, and—as is any mode of caring—self-affirming

3. "On Caring," p. 166.

4. "Love as Valuing a Relationship," p. 142.

concern for the existence and the good of what is loved." Since the "lover's concern is rigidly focused in that there can be no equivalent substitute for its object, which he loves in its sheer particularity and not as an exemplar of some more general type," loving cannot be "the rationally determined outcome of even an implicit deliberative or evaluative process."[5] But this account of irreplaceability comes at an unpalatably high cost. Ellen's catechism illustrates how natural it is to give and ask for reasons for love. Catherine's initial refusal to give a reason feels like a dodge. The problem with her later answers is that the reasons she gives are bad ones, not that she is making a category mistake in giving them at all.

Two established lines of criticism underscore this point. First, it is simply not plausible that love consists in the attitudes Frankfurt claims it does. There is a difference between loving someone and assuming, for no further reason, the project of being the agent of someone's interest. Velleman observes that at "the thought of a close friend, my heart doesn't fill with an urge to do something for him, though it may fill with love."[6] I care about my close friends, and would do a lot for them if they asked me (and in some cases even if they didn't). But there are a lot of helpful things I could do for my friends that I feel absolutely no desire to do: their laundry, for instance, or their grocery shopping. It's not that such desires are overridden by others, or that my friends would find it off-putting if I acted on them. (Even if they insisted that they would *not* feel infantilized if I did their laundry for them, this would not incline me to do it.) It's rather that these sorts of things just aren't what friendship is about.

Second, there are some things it just doesn't make sense to love.[7] Suppose you were gripped by an involuntary, nonutilitarian, rigidly focused, and self-affirming concern for the

5. *Taking Ourselves Seriously and Getting it Right*, pp. 40-41.

6. "Love as a Moral Emotion," p. 353.

7. This point comes from Troy Jollimore, in *Love's Vision*, p. 22-23.

existence and the good of a random picnic table.[8] One weekend you eat a hot dog there, but on the drive home your thoughts keep returning to it, a vague fondness rising in your breast. So every weekend thereafter you go back to Table 7-G to clean it off, protect it from the elements, replace rotting beams, and so on. This attitude is not just unusual. It's positively perverse. Love for a random picnic table is either irrational or unintelligible. We need to explain why, and we need reasons for love to do it.

3 Loving someone as a rational agent

While Velleman agrees with Frankfurt about the "ineluctably particular" nature of a beloved's importance, he denies that it prevents love from being a rational response to a generally-held valuable property. In fact, he argues, it's a response to a valuable property that all persons share by definition: one's bare rational nature—a property whose value must be appreciated in ways particular to each instance of it.

In the Kantian framework Velleman assumes, your rational nature is what makes you worthy of being valued as you are in yourself. The responses it warrants come in two varieties: respect and love, "the required minimum and optional maximum responses to one and the same value."[9] Velleman attributes the special character of each of these attitudes to the fact that the value of one's rational nature is independent of, because prior to, the value of any properties that not all persons share equally—and hence which distinguish particular persons from one another. This is why respect is equally owed to everyone, and why it consists, roughly, in according each individual the basic regard to which one's dignity as a person entitles one. The same rational independence explains the special *openness* appropriate to the individual value of a beloved.

8. The reader who doubts that a picnic table can have a good may substitute a patch of grass, or small animal.

9. "Love as a Moral Emotion," p. 366.

Whereas the demands of respect are predominantly negative—they consist, primarily, in prohibitions against treating people in ways that ignore their value as persons (for example, by manipulating or exploiting them)—love involves a positive, arresting awareness of that value.

More specifically, love "arrests our tendencies toward emotional self-protection from another person, tendencies to draw ourselves in and close ourselves off from being affected by him. Love disarms our emotional defenses; it makes us vulnerable to the other."[10] This means that it isn't to be identified with any specific motives or emotions, but rather with a heightened sensitivity to the significance of whatever specific characteristics, attitudes, or interests a beloved manifests: for Velleman, "a sense of wonder at the vividly perceived reality of another person"[11] is the closest thing to a constitutive feeling of love there is. Thus, love exposes you to a very wide range of emotional responses—not all favorable—corresponding to the wide range of features a beloved might exhibit on a given occasion. You might be thrilled by the admiration of your beloveds, hurt by their insults, and solicitous of their needs, even when the same admiration, insults, or needs would barely register if observed in casual acquaintances.

Though Velleman sometimes motivates this feature of love on phenomenological grounds, he recognizes that its importance goes deeper. It captures, he argues, the way people value their beloveds as special. Being valued as special "doesn't entail being compared favorably with others; it rather entails being seen to have a value that forbids comparison. Your singular value as a person is not a value that you are singular in possessing; it's rather a value that entitles you to be appreciated singularly, in and by yourself."[12] This implies, minimally, that if you really love someone, you must be open to unanticipated developments in your own ends in light of

10. ibid., p. 361.

11. "Beyond Price," p. 199.

12. "Love as a Moral Emotion," p. 370. For another forceful defense of this requirement, also on Kantian grounds, see Rae Langton, "Love and Solipsism."

what that person has to show you. Loving people means taking them as they come. Your attentiveness and vulnerability to them can't be contingent on how well they serve your existing ends or conform to your prior ideals, and hence you can't suppose yourself to need any special reason for your heightened sensitivity to whatever is significant to or about them. As an illustration, Velleman describes watching his sons grow up:

> In a quick succession of years I became deeply interested in lacrosse and Morris dancing, poetry slams and photography, and specifically in the accomplishments of a particular midfielder, Morris dancer, poet, or photographer, because these were the directions that my children had set for themselves. Of course, I eventually learned to appreciate some of these accomplishments intrinsically: I would realize with amazement that I was cheering as my son walloped a schoolmate with a metal stick or that I was applauding choreography that previously would have struck me as no more than quaint. But I learned to appreciate these accomplishments, to begin with, because they were the ones that my children had chosen to cultivate.[13]

But while Velleman is correct to stress the characteristic openness of love, I think he gets its details wrong, because he misconceives the kind of value to which love for an individual is responsive. People value their beloveds as incomparable, but only up to a point. They still take themselves to have reason to love some people *and not others*.

On Velleman's gloss Catherine would be perfectly correct to love Heathcliff for his similarly-constituted soul. Their "souls" are their bare rational natures, and these are indeed exactly alike. Her mistake is just in viewing *Linton's* soul as any different. Catherine's disposition to be vulnerable to Heathcliff's rational nature (which presumably constitutes her love for him)[14] is a strictly causal matter, an incidental quirk of her psychology. But clearly that's not how she loves Heathcliff, or how she should. Catherine doesn't see him as someone she just *happens* to be

13. "Beyond Price," pp. 205-206.

14. Velleman is not explicit about the kind of psychological state in which love consists, and much in his presentation can instead suggest a view of it as an *occurent* state of arresting awareness. The latter interpretation, however, should be rejected on grounds of charity. You no more cease to love someone when you are vexed or preoccupied than you forget what they look like when you close your eyes. If this weren't so, then either most people would love much less than they think they do, or love would be even more of a headache than it already is. (This point comes from Susan Wolf's first Phi Beta Kappa Romanell lecture.)

arrestingly aware of, like someone you happen to be standing next to at a party and might as well make small talk with. She *endorses* her love for him, specifically.

Consider that the moral and prudential considerations in favor of Catherine's loving Linton instead are substantial. If she could somehow replace her disposition to arresting awareness of Heathcliff with one directed toward Linton, it would help her keep a promise, ease her into a life of comfort and prominence, and orient her toward safer, more socially acceptable, and morally improving pursuits. Given that, on Velleman's view, Catherine has no special reason to love Heathcliff, why shouldn't she regard her disposition to arresting awareness of him as other than a mere inconvenience? Of course, people don't regard their loves as dispositions to be managed at their moral and prudential convenience. For Catherine, suddenly ceasing to love Heathcliff is unthinkable. The prospect would appear as a disturbing failure to appreciate his profound significance to her. In short, it is essential to her love that she experience it as more than merely optional.[15]

One might, therefore, ask why Velleman shouldn't just jettison his claim that love is rationally optional. It will be instructive to consider this possibility. So modified, Velleman's view would place love on the same level with respect, as a rationally required response to rational nature as such. Just as you have reason to respect everyone, it would hold, so too do you have reason to love them. It's just that it's generally much harder to love people than it is to merely respect them, and nobody is in a position to blame you for failing.[16] So those of us who fail to love as we rationally ought, even those of us who fail radically—presumably, more or less all of us—do so forgivably. Velleman clearly aims to avoid this view, and it is easy to see why. It

15. Or, as Catherine herself puts it: "My love for Linton is like the foliage in the woods. Time will change it, I'm well aware, as winter changes the trees—my love for Heathcliff resembles the eternal rocks beneath—a source of little visible delight, but necessary."

16. Or at least, on the religious version of this view, nobody on earth.

completely abandons the idea that love may be unapologetically selective in any but the most superficial sense. But I have just argued that the degree of selectivity Velleman is actually entitled to is pretty superficial anyway. The difference is just that on the present modification one is to regard one's psychological inability to perfectly love everyone as a genuine rational imperfection, one we have reason to work to overcome.[17]

This answer is very honest in its way, and should not be dismissed out of hand. The idea that we are all to love one another unconditionally is historically important and very powerful. It is fundamental to the Christian ethical tradition, and may, not unrelatedly, be truest to the Kantian spirit. So perhaps it should not be surprising if this is really where Velleman's theory leads him. (I suspect it is also the only coherent conception of love as a truly *moral* emotion.) But the resulting ideal of love is not only demanding, but strangely impersonal. It collapses, or at least trivializes, the distinction between love for particular persons and love for humanity as such.

4 Loving someone as a relative

Love for *particular* persons, therefore, must be a *selective* rational response to valuable properties of the beloved. But not just any sort of property will do. Linton's attractive qualities, remember, would make him too replaceable. Here, Niko Kolodny's appeal to *relational* properties represents an important advance.

Kolodny argues that reasons for love are grounded in the value of one's ongoing historical relationships to one's beloveds. I love my brother, for instance, because *he's my brother*. We have the same parents and we grew up together. This makes it easy to explain irreplaceability: it is important to me that I love my brother, in particular, because he is the particular person to

17. Perhaps Velleman feels the influence of this point in his suggestion that we are not equally inclined to love everyone in part because the "human body and human behavior are imperfect expressions of personhood, and we are imperfect interpreters." ("Love as a Moral Emotion," p. 372)

whom I stand in the fraternal relation. Other people might stand in the fraternal relation to *someone*—they might be other people's brothers—but that doesn't give me reason to love them. They don't stand in the fraternal relation to *me*. (Now, it so happens that I have two brothers. Kolodny's view implies that I therefore have just as much reason to love the one as the other. But this is just as it should be. I do have reason to love both equally.)

More specifically, Kolodny argues that relationships of certain types are *finally* (that is, non-derivatively) valuable, and so constitute sources of reasons for love. For him, loving someone consists in (i) believing your relationship to that person to be an instance of a valuable type, and thus (ii) taking it to be a reason both for being emotionally vulnerable to, and for acting in the interest of, both your beloved and the relationship itself, in ways appropriate to relationships of that type, and (iii) believing that others in relationships of the same type would have similar reasons for similar attitudes concerning their own beloveds.[18] When you love someone, then, you value both your beloved and your relationship itself—but it's the relationship, in virtue of its general type, that you see as the source of your reasons to accord special value to each. This isn't to say, of course, that lovers don't have reason to be interested in whatever valuable qualities their beloveds might have. On the contrary, my love for my brothers involves, in part, my taking a special interest in all kinds of great things about them. But on Kolodny's theory, my reasons for doing so aren't grounded in the value of the qualities themselves—*those* reasons would apply to anyone—but in the value of my relationships to their particular bearers.

It's technically conceivable that my fraternal relationships could have provided me with special reason to care about my brothers even if they were only derivatively valuable, since I still wouldn't have them to anyone else. But by construing their value as final, Kolodny elegantly accounts for the thought that love should be *constant*: its reasons should endure through a suitably

18. This is a compressed version of Kolodny's formulation in "Love as Valuing a Relationship," pp. 150-151.

wide range of changes in a beloved. I leave "suitably wide" vague on purpose, but the basic idea should be intuitive. If my brothers came to have very different valuable qualities, I'd have equally good reason to take a special interest in those. I'd even have reason to care about my brothers in the (distant!) possible world where they didn't have much going for them at all. For Kolodny, the explanation for this is simple: so long as your beloved remains such that your relationship can somehow endure as a valuable instance of its type, your reasons for love endure as well.[19]

Like Velleman's, Kolodny's theory captures a genuine and valuable form of love. We've already seen how plausibly it describes familial love; it likewise accounts for the possibility of loving attachment to things other than persons, like pets, gardens, cars, or institutions.[20] Further, I think Kolodny is right that our histories with our beloveds are ultimately what make them irreplaceable to us. But they can't always do so in the way he thinks. For some of the deepest loving relationships *can't* be finally valuable: the value of your beloved must come first, and your relationship must be valuable to you only as a relationship to the particular person in question.

As a first effort at bringing out what Kolodny's view misses, observe that a surprisingly wide range of relationships count as loving by his criteria. One example is teaching. It goes without saying that teachers who value their pedagogical relationships take them to be reasons for acting in the interests of their students in pedagogically appropriate ways. And while good teaching probably does not require overbearing, *Dead Poets Society*-style sentimentalism, it is plausible that the best teachers, and the ones who get the most out of it, further take their pedagogical relationships to be reasons for certain forms of emotional vulnerability to their students—for taking pleasure in their students' progress, for instance, and being troubled by their

19. Of course, relationships can end. Waning sexual attraction generally ends romantic relationships, or turns them into friendships; even familial relationships may fail to survive sufficiently gross betrayals (as numerous movies and television shows about mobsters memorably attest).

20. While Kolodny explicitly defines "relationships" in his sense as essentially interpersonal, I see no reason to view this limitation as more than stipulative.

unnecessary confusion—and for corresponding actions and attitudes toward their pedagogical relationships themselves. For example, they are unsatisfied with pedagogical relationships that are going badly or that end prematurely—for example, with their students dropping out—and when appropriate they act to prevent these things.

Now, it is arguably possible to love someone specifically as a student. But it would be a stretch to say that if you value your pedagogical relationships in this way, you necessarily love every student with whom you have one. What you love in this case is *teaching*. The students you love, if any, are the special ones, who "make it all worthwhile." They are the ones with whom you value your pedagogical relationships not just as instances of a generally valuable class, but additionally as pedagogical relationships that are especially valuable *because of specific characteristics of the student in question*. Thus, Minerva might spend an extra hour helping Neville, who is pleasant enough and tries hard, out of her love for teaching, and yet do the same for Hermione, who is brilliant and delightful, additionally out of pedagogical love for *her*. It will be a matter of indifference to her that her pedagogical relationship is to Neville, as opposed to any of the many other adequate Hogwarts students whom she might have taught but happened not to. But not so with Hermione. Yet it is exactly this distinction that Kolodny's view lacks the resources to draw.

It might be offered on Kolodny's behalf that while Minerva may indeed have a special relationship to Hermione, that relationship could be a finally valuable relationship of a different type. Minerva could, for instance, simply be Hermione's friend.[21] But once we see how the distinction between finally valuable relationships and those whose value depends on the specific person in question pertains to teaching, its relevance to more intimate cases becomes apparent.

21. It might also be suggested that the difference in value is merely one of degree. Hermione's virtues might make a difference simply by enabling an especially valuable instance of the same type of relationship Minerva has to Neville, not (as I have supposed) by adding an extra *dimension* of value specific to Minerva's relationship to *Hermione*. But whether or not this is plausible for teaching, it is not for marriage.

It's possible to value a marriage—even deeply and for its own sake—to someone you don't love. This may be how Alexei Karenin viewed his marriage, for instance, at least before he learned of Anna's infidelity. As Anna's husband, Karenin would indeed have reason to take a special interest in Anna's attractive qualities, like her sensitivity and verve—but only because those were the qualities that happened to be instantiated in his wife. There'd be nothing about Anna's sensitivity and verve as such, much less about Anna herself, that gave *those particular qualities* a special claim on Karenin's attention. Had Karenin been married to someone else, *her* qualities would have been just as lovable to him; Anna's would have been merely attractive.

Kolodny claims, in response, that "it doesn't seem like a distortion to say…that a wife wants to be loved by her husband, at the deepest level, because she is the woman with whom he fell in love and made his life. […] Let us suppose that they had never met and had made their lives with other people. Imagining herself in that situation, would she still want him to love her? Would it make sense to her if he did?"[22] At first glance, this reply seems right. But I think this is only because it subtly trades on exactly the point Kolodny's critic should press against him. If love is indeed grounded in the value that certain characteristics of your beloved have for you, it is plausible that the identity of those characteristics depends on the sort of person you are. But love changes you. Had the spouses in Kolodny's example married different people, they would have become different sorts of people themselves. Of course they would have found different characteristics lovable.

In order for Kolodny's example to vindicate his theory, therefore, we need to hold fixed[23] the nonrelational properties of all parties concerned. Suppose you are married to a person named

22. "Love as Valuing a Relationship," p. 157.

23. Or, at least, as fixed as possible. It can be hard to draw a sharp distinction between putatively nonrelational properties of someone's personality and character from relational ones like memories and ongoing concrete projects. This theme will turn out to be important later, but bracket it for now.

Smith. Your marriage is a happy one; you and Smith delight in and admire all kinds of things about each other. You also happen to have a colleague, Jones, whom you dated for a semester back in college. If you gave the matter much thought, you'd admit to yourself that you'd likely have grown to love each other if you stayed together, and even that you still find her fairly attractive. But you really don't give the matter much thought. Jones is nice enough, but you can't imagine how anything about her could grip you the way so much about Smith does. As far as you're concerned there's no contest. It's crazy to think that, married to Jones while remaining exactly the sort of person you are now, and knowing as much about the two as you do, Jones would be just as lovable to you as Smith now is, and Smith's radiance would be dimmed to that of a moderately attractive acquaintance. But that implies that your reasons for loving Smith do not derive, ultimately, from valuable properties of your marriage to him. Rather, they derive from valuable properties of Smith himself.

5 Loving someone for a self like yours

Let me stress again that the fond attachment Kolodny describes, like the unconditional loving awareness Velleman does, really does constitute a legitimate and admirable form of love. But the shortcomings of both their views show that we also need a third form, grounded in a selective appreciation of a beloved's distinguishing features. This, I will argue, is the kind of love Catherine has for Heathcliff.

When Catherine says she loves Heathcliff for what his soul is made of, I assume that she is referring to his *identity* or *character* (I'll use these terms interchangeably), as defined by the values by which he finds it fundamentally worthwhile to live. Now, I haven't yet done anything to show why the properties that constitute selective reasons for love must, specifically, be the beloved's values. In fact, it might seem perfectly normal to love someone for other things. Wealth might be

mercenary, and handsomeness superficial, but what's wrong with loving people for their cheerfulness—or, for that matter, for their intelligence, sensitivity, or senses of humor?

The enduring popularity of *Wuthering Heights* itself suggests, however, that there is something about Catherine's kind of love that many people find compelling. At least in rough outline, it's easy to see why. To begin with, it seems on reflection that whether someone finds your other qualities *lovable* (rather than just interesting, sexy, or otherwise pleasing) normally does depend, a great deal, on what those qualities say about your values—on the interests or drives animating your intelligence, say, or the outlook on life embodied by your jokes, or even the sensibility exhibited in how you dress or walk.[24] Further, and more basically, someone who loves you selectively, for the specific values you identify with, sees you for who you distinctively are—and finds you distinctively valuable as such. There's something wonderfully affirming and empowering about this—especially when you love the other person in the same way—though it's hard to explain what. (I'll offer more of an explanation below, once I've said more about what Catherine's kind of love involves.)[25]

Once we've seen how essential irreplaceability, openness, and constancy are to love, however, it can seem strange—if not paradoxical—that the idea of being loved for your values could be so appealing. For while these features pose a challenge to any conception of love as a form of selective appreciation, the challenge seems especially severe in the present case. Start with irreplaceability: while the values Catherine loves in Heathcliff might be *unusual*, there's no reason to think they're essentially *unique* to him, as would seem necessary for it to be important to her that she love him in particular. On the contrary, they can't be, if Catherine is correct in

24. Martha Nussbaum makes the same observation, in defense of a similar thesis, in "Love and the Individual: Romantic Rightness and Platonic Aspiration," in her *Love's Knowledge*, p. 327.

25. For another attempt at an explanation, to which I am substantially indebted, see C. S. Lewis's beautiful and trenchant discussion of friendship in *The Four Loves*, ch. 4.

taking herself to share them. And it's with respect to the beloved's values that openness and constancy seem most important. Think of lovers who *aren't* open to their beloveds as they are in themselves—who respond to any deviations in their beloveds' values from their own with disappointment, rather than welcoming their beloveds' novel concerns as potential enlargements of their own perspectives. Such love—if it's intelligible as such at all—seems narcissistic on the part of the lover, and insulting to the beloved. And if these putative lovers further saw the *continuation* of their relationships as rationally contingent on ongoing adherence to the party line, their interest would go from a distressing narcissism to a downright sinister possessiveness.

What I want to argue now, however, is that the idea of loving people for values they share with you only faces these problems if we think of the values in question themselves as fixed, static things. I don't think values are like this—not all of them, at any rate. I think instead that a beloved's values are loved as things that are essentially in the process of being determined, through an ongoing sequence of judgments and actions. To that end, I'll defend this conception of agency in the remaining sections, showing how it yields an account of the nature and value of love for persons as particular individuals that vindicates both Ellen's catechism and Catherine's ultimate answer to it.

6 Improvising with a partner

In the last chapter, I appealed to musical improvisation to model how agents can act for reasons they're making up as they go, just as musicians can literally work out the musical ideas they are trying to express through the processes of expressing them.[26] Now we can extend the model to multiple agents. To do so, let's go back to Keith Richards. He reports:

26. In *The Retrieval of Ethics*, Talbot Brewer makes a similar appeal to musical improvisation as a paradigm of what he calls "dialectical activity," which he likewise conceives as a process through which agents refine their conceptions of their ends through their efforts at achieving them. However, Brewer's theory differs from mine in that

> There's something beautifully friendly and elevating about playing music with people. This wonderful little world is unassailable. It's really teamwork, one guy supporting the other, and it's all for one purpose, and there's no flies in the ointment, for a while. And nobody conducting, it's all up to you. It's really jazz—that's the big secret. Rock and roll ain't nothing but jazz with a hard backbeat.[27]

What makes this so great, Richards insists, is the experience of pursuing a shared end:

> You're sitting with some guys and you're playing and you go "Ooh, yeah!" That feeling is worth more than anything. There's a certain moment where you realize you just left the planet for a bit and that nobody can touch you. You're elevated because you're with a bunch of guys that want to do the same thing as you. And when it works, baby, you've got wings. You know you've been somewhere most people will never get; you've been to a special place. And then you want to keep going back and keep landing again, and when you land you get busted. But you always want to get back there. It's flying without a license.[28]

Contrast these remarks with Richards's account of improvisation in the first person from the previous chapter. There, he plainly did not have a determinate end in view. Yet here he is elevated precisely because he views his bandmates as playing with a deep sense of common purpose—as sharing his end. This may seem paradoxical. In what sense could it be intelligible for Richards to take his bandmates to have the *same* end as he, without implicitly presupposing there to be a determinate fact of the matter as to what the contents of their respective ends are?

The answer is that when you take other improvisers to share your end, just as when you take your own end to call for a certain action, you aren't making a judgment about the content of your end as it stands. Richards's sense of his bandmates as "wanting to do the same thing as [he]" is best interpreted, I think, as analogous to the feeling, in the context of individual improvisation, that a certain riff is somehow the way to go at the time. It registers the acceptance

---

it is explicitly and unapologetically Platonic. Whereas I use improvisation to model a process through which agents freely create their ends for themselves, Brewer argues that agents gradually acquaint themselves with the ideal forms in which activities of certain types are to be pursued. This leads to some strange results. Jazz turns out to be an effort to apprehend and instantiate objective aesthetic ideals, rather than an act of personal expression. And while persons are properly loved for their developing evaluative outlooks, an "evaluative outlook is properly loved only because and to the extent that it exemplifies the zeal for adherence to objective truths about the good that is the proper *telos* of the human capacity for practical reason." (p. 256) So much for Catherine and Heathcliff.

27. *Life*, p. 105.

28. ibid., p. 97. Richards makes this observation while recounting an early gig that included Mick Jagger and Brian Jones, but neither Bill Wyman nor Charlie Watts. For this reason I hesitate to refer to "Keith Richards and his bandmates" by the obvious proper noun.

of a commitment, rather than a judgment about what, normatively speaking, is already determinately the case. Here, however, Richards's commitment doesn't concern something he takes himself to have reason to do on a single occasion, but rather the things his bandmates take themselves to have reason to do over time. He commits, that is, to treating what *they* take themselves to have reason to play, given what *they're* trying to express, as standing in the same mutual explanatory relation to what he has reason to do that his *own* responses do—they're things he'd have reason to play himself, had he been in their position. Thus, Richards's attitude is analogous to that someone who is trying to adhere to a certain set of standards, and who, while partially ignorant of the content of those standards himself, does know of someone who is also trying to adhere to them, and doing so more or less successfully. In effect, then, he accords his bandmates a presumption of default authority with respect to his end.

This presumption of normative authority is, I think, crucial to the nature and value of loving relationships.[29] Its significance, however, is analogously reflected in attitudes among musical improvisers, as Richards's case illustrates. This is manifested, most basically, in the complex attitude of open-ended receptivity and sense of individual purpose Richards might plausibly be viewed as taking toward his bandmates. So interpreted, what Richards and his bandmates all want to do is express the musical ideas they're vaguely reaching for; in playing together, they recognize their respective ideas to be the same. So when Richards hears what his bandmates are playing, and he's moved by the feeling behind it, he's moved specifically by something that feels expressive of precisely what he's reaching for himself. It's easy to see how that could be "beautifully friendly and elevating." If you're improvising with someone, and your partner responds with something that complements your playing so well it feels like a fuller

29. Versions of this claim are not unfamiliar in the literature. For two notable recent ones, see Elijah Millgram, *Practical Induction*, ch. 7, and Kyla Ebels-Duggan, "Against Beneficence," esp. pp. 158-159.

expression of the musical idea behind it, it feels affirming, and enhancing. The thought is something like: "So *that's* how to do it!—that's what was cool about where I've been going. And better still, now I know to play *this*"—and you respond reciprocally to your partner. Yet when you do respond by trying to play in a way that coheres with your partner's playing, you're not just returning the favor. You're doing exactly what feels like the natural next stage in your own musical project. The result is a feedback loop of mutual, spontaneous exchange: you express what feels right to you; your partners are spontaneously moved by what they feel in your playing to play what feels right to them; you experience their responses as an apt development of what you're playing and delightedly reply to it as such; you're off to the races.

7 Loving someone as a partner in deep improvisation

When Richards and his bandmates recognize and respond to each other as pursuing the same end, we can say that they value one another as *improvisational partners*. In this, they share a relationship with the same general structure as relationships between lovers. To love someone in particular, I submit, is to value that person as a partner in deep improvisation.

An improvisational partnership is a type of ongoing relationship grounded in the partners' mutual recognition of one another as sharing an end with respect to a given activity.[30] In principle, any improvisational activity can be done with a partner: partners in musical improvisation explore a common musical idea in their playing; students may improvise with each other as partners in working out an interpretive approach to a text; spouses may improvise with each other as partners in working out the terms of a marriage. None of these relationships necessarily involve partnership in deep improvisation (in which the partners would work out together the basic significance that their activities of musical expression, textual interpretation, or

30. Or, equivalently, with respect to a cluster of interlocking activities (which need not be clearly delineated).

marriage had in their lives), but all of them can. As a special case, partnership in deep improvisation can also be *global*, as it is for Catherine and Heathcliff, such that the partners work out their entire approach to life together. But it need not be. I love my grad school officemate for much of what he sees in philosophy, but not all of it, and while we have a very satisfying friendship, we each think the other's political views are pretty awful.

Like relatives in Kolodny's sense, improvisational partners value their partnership itself as well as one another. But unlike relatives in Kolodny's sense, they do not value their partnership *finally*. Rather, they value it only because and insofar as they recognize one another as warranting a presumption of authority in judgment with respect to the relevant activity—that is, only because they are justified in taking one another to share their respective ends.

It follows that there are selective reasons for love (and for improvisational partnerships in general), but reasons of a special kind. Here, again, taking yourself to have reason to value someone as an improvisational partner is analogous to taking yourself, in the individual case, to have reason to perform a certain action. In neither case does it imply that the action or attitude in question was rationally determined in advance: had you not taken yourself to have reason to value someone as a partner you probably would have been right, since you'd have developed your ends in a different direction. This, I think, is why the thought that love is arational or rationally optional can seem so natural: it's deceptively close to the truth. But, as I argued above, there's a crucial difference between recognizing that a response wasn't rationally determined in advance, and regarding it as arational or rationally optional *having actually made it*. When you improvise, you do the former, but not the latter: you regard your decision to respond in one way rather than another as something that admits of and requires—and might not get—rational justification, in terms of the end you're thereby determining yourself. This is, again, compatible with the thought that had you done the other thing, you might, too, have had reason to do it—

just reason relative to a different end. This makes the authority that Richards accords his bandmates more than just epistemic. For, having accorded his bandmates this authority, the ways he takes himself to have reason to play will depend on the ways his bandmates do—and so, therefore, will the content of his end itself. This in itself is crucial for love, because it is what explains how people who love each other for identifying with the same values do not just *mirror* each other's values but *shape* them.

So valuing someone as an improvisational partner is, if things go well, a self-fulfilling prophecy. That's why it's perfectly intelligible to do so without supposing it to be possible—even in principle—to provide a justifying explanation for this decision at first, in light of the similarity of your respective ends. For you may rather suppose yourself to be in the process of working out that explanation, by improvising with that person and seeing what the two of you turn out to be like, and what, in light it is, just what it is about the other person that so resonates with you. (Of course, things might not go well: you might fail to find anything in your partner that helps you make sense of your initial attraction. If this happens, your attitude will turn out in retrospect to have been unjustified.)

This may seem strange, but in connection to love it's really the most obvious thing in the world. The language of love is full of terms for inexplicable but warranted attraction: there's electricity, that twinkle in his eye, that special something, that *je ne sais quoi*. (The language of music is not dissimilar in this respect.) We've all encountered or heard of people who admit it to be impossible for them to describe their reasons for loving the specific people they do (or, at any rate, to describe them any more clearly than Catherine or Montaigne did), but nevertheless insist that they do have such reasons. I imagine most of us philosophers have felt the temptation to quietly conclude that such sentiments, while romantic, betray a basic confusion about rationality.

But if I'm right, the romantics have been right all along about an important class of reasons that we, for the most part, have missed.

* * *

Interpersonal love differs from other forms of improvisational partnership because the ends lovers share constitute fundamental values with which each identifies. This makes the interest and responsiveness warranted by the similarity of their ends correspondingly more profound. To flesh out what these attitudes involve, let's return once more to Catherine and Heathcliff.

Catherine and Heathcliff love each other for their common wildness. When she tries to explain this to Ellen, Catherine recalls a dream of going to heaven but being miserable there. Ellen points out, reasonably enough, that of course she wouldn't like it: heaven isn't supposed to be the sort of thing sinners would like. "This is nothing," Catherine retorts.

> "I was only going to say that heaven did not seem to be my home; and I broke my heart with weeping to come back to earth; and the angels were so angry that they flung me out, into the middle of the heath at the top of Wuthering Heights, where I woke sobbing for joy. That will do to explain my secret, as well as the other."

Wuthering Heights is a home to Catherine because of the free and vigorous way of life its rawness, beauty, and isolation enables her to lead there. Unlike anywhere else in her life, it provides opportunities for creative exploration and discovery, physically robust activity, and uninhibited emotional expression—all things to be approached in a very different spirit than the domesticated concerns in which she and Heathcliff are otherwise expected to participate. As improvisers, Catherine and Heathcliff are each engaged in a process of working out just what this wildness means to them: what precisely is to be appreciated in being in the wilderness, and how—and how the spirit of wildness each prizes is to be embodied in an overall approach to life.

By sympathetically engaging with what the other sees, Catherine and Heathcliff offer each other focus and reinforcement. Catherine's judgments and actions serve as a guide for

Heathcliff. If something seems worth doing to her, he'll see, and feel, this as a hint about what he himself has reason to do, and respond accordingly. I'd imagine most of these instances of shared practical reasoning are small and subtle. They might concern things like what's to be savored in an autumn wind, or what's interesting about a certain bird, or how and why the curate is to be tormented today. In loving Catherine as a partner in deep improvisation, Heathcliff will be drawn to her approach to life. In viewing her judgments as warranting a presumption of normative authority, Heathcliff will experience them as attractive, as having a rightful power to shape his own sense of his values so as to accord with them. Similarly, in seeing what Catherine takes to be worthwhile as to be taken into account in his own normative explanations, Heathcliff will experience her personality as calling out for attention and understanding—that is, he'll find it fascinating. Over time, these instances of sympathetic engagement and exchange add up. They enable the lovers to determine and act from more richly illustrated conceptions of value than otherwise would have been available to them.

It is this mutual self-creation that explains the irreplaceability, openness, and constancy characteristic of interpersonal love. I'll address these features in opposite order to that in which they were introduced. Conveniently, this turns out to be in ascending order of complexity.

Constancy: That partners in deep improvisation have reason to continue to love each other through a wide range of developments in one another's values, not necessarily capable of being anticipated in advance, should be obvious by this point. Someone who loves you as a partner in deep improvisation loves you, throughout your relationship together, for a specific set of values with which you identify. Because the content of these values is in the process of being determined, however, the reasons for love they constitute endure as their content is continually being reshaped and refined. Change is not the exception but the rule.

Remember that even Sonnet 116—probably the single most quoted paean to constancy in English—begins by describing love as the marriage of true minds. This suggests two pertinent observations. First, the fact that reasons for love are constituted by a person's values, rather than external characteristics like Linton's attractions, itself means that love for a partner in deep improvisation can be expected to survive the sort of surface changes it really, obviously should. Second, constancy is important, but so is discernment. Of course it can be appropriate to cease to love a person, if one or the other of you undergoes a fundamental change in character or if the two of you do not turn out to share as much as you thought.

Openness: Partners in deep improvisation are open to each other as they are in themselves because of the distinctive way the values they share are shaped by their particular interactions. As I explained above, that an improvisational partner takes some action to be appropriate is in and of itself a *prima facie* reason for you to do so as well, in virtue of the presumption of authority appropriate to a partner as such. And because the actions you thus take to be appropriate determine the content of your ends, the bare fact that your beloved responds to a particular case in a certain way can in and of itself make a difference to your values. Thus Velleman was right to stress that you don't need any special reason for heightened sensitivity to whatever is significant to or about your beloved—in any given case, that you love the person is reason enough.

Recall that both Frankfurt and Velleman took love for you as you are in yourself to necessarily not admit of selective criteria. Clearly, someone who loved you only because and insofar as you satisfied certain antecedently fixed criteria ("handsome, check; young, check; cheerful, check; rich, check!") would not really be loving you for who you are. We can now see that while Frankfurt and Velleman correctly identified a problem with viewing love as rationally subject to selective criteria, they misidentified its source. It is fine for love to be subject to selective criteria—in fact, the inadequacies of Frankfurt's and Velleman's views have shown that it must

be. What is not fine is for it to be subject to *antecedently fixed* selective criteria. The improvisational model shows the difference between the two. When you love someone as a partner in deep improvisation, you love that person according to criteria that are not only in principle impossible for you to articulate in advance, but which depend on your beloved as much as on you. Love for people in their sheer particularity is not love that doesn't admit of reasons; it is love that requires you to respond to beloveds in their sheer particularity to determine the reasons you have for it.

Irreplaceability: The most basic reason why partners in deep improvisation are irreplaceable is simply that they are incomparable: the nature of the values in question makes the possibility of a replacement incoherent. Someone counts as a suitable replacement for an improvisational partner if that person enables you to realize the same value in the relevant activity that the original did. This may be possible in most forms of improvisation: if you're ultimately in it for the money or the adulation, one bandmate may be just as good as another, even though you'd be expressing different things with each. But it is not possible in deep improvisation, since the value of the activities you share is itself something your partner plays an ongoing role in determining. So any standards by which putative replacements might be assessed are epistemically and ontologically posterior to continued engagement with the original. If you had a different partner, you'd have different standards: there's no common basis of comparison.

Now, anyone you take to share values with which you identify will be incomparably valuable to you in this way, even when the interest isn't mutual. But when it is mutual, lovers become irreplaceably valuable to each other in a deeper sense. Catherine says of Heathcliff that he "comprehends in his own person my feelings to Edgar and myself."[31] Interpreted as an

31. Catherine's statement occurs at the beginning of a speech that has rightly worried many critics. She goes on to proclaim that all her miseries in life have been for Heathcliff, that he is her great thought in living, that the world would be empty without him, and so forth. On my view, the selflessness Catherine expresses is incidental to love proper. Heathcliff is properly lovable to Catherine because he helps her live as more fully herself, not because he gives her something to live for. (It helps to remember here that she speaks as a moody and theatrical fifteen-year-old.)

improviser, she is referring to Heathcliff's access to the values with which she identifies: as her lover, he can access them in ways that nobody else, in principle, can. He can interpret and develop them through the lens of his own history while still appreciating them as she does.

Catherine works out who she is by improvising from an evaluative currency of things like interesting birds, invigorating autumn winds, and obnoxious curates. She takes these things to be significant in ways that are to help explain how it will be important to her to live going forward, and that are to be explained, in turn, by their relation to the overall way of life they help constitute. This makes her understanding of herself and her values essentially historical, and particular. The only way for anyone else to understand them from her perspective is to attribute the same practical significance to particular cases that she does—to see, in the same way that she does, those cases as contributing to, and helping to explain the nature and attractiveness of, the kind of life it is important to her to lead. But I have just argued that to be committed to attributing the same significance to particular cases that you do is precisely what it is to love you. So Catherine and Heathcliff understand each other the same way they understand themselves: through a joint history of particular interactions, constituting a common evaluative currency.

But while any number of lovers might share your sense of your identity, just as any number of bandmates might share your musical idea, Heathcliff's *perspective* toward Catherine's values is all his own—as, crucially, is his role in enabling her to develop them. In loving Catherine as a partner, Heathcliff integrates what she has shared with his own evaluative history; in acting according to the conception of his values he thereby forms, he may inspire Catherine, in turn, to develop her own values in new directions.[32] (Recall Velleman's discussion of his sons from §3.) But the contributions he has to offer would necessarily differ from those of a different

32. In stressing the essentially historical character of the roles lovers play in drawing one another out, I follow Amelie Oksenberg Rorty, "The Historicity of Psychological Attitudes," and Alexander Nehamas, "The Good of Friendship," among others.

lover, with a different history. Even if Heathcliff disappeared from Catherine's life, and she eventually fell in love with someone else, her new lover would not understand her in quite the same way as Heathcliff did, from quite the same point of view. So long as she continued to identify with the values she and Heathcliff worked out together, there would remain a side of her accessible to Heathcliff alone, things he alone could bring out of her.

Why, then, might it be so important to us to share ourselves with our lovers? Let me finish by sketching the beginning of an answer. To begin with, the things our lovers bring out of us might not be things we're capable of bringing out ourselves. This is underscored by the fact that the best lovers often seem like opposites: consider Elinor and Marianne, Holmes and Watson, Kirk and Spock. These people all share certain fundamental concerns with their partners—ones centered, respectively, around ideals of feminine autonomy and the enjoyment of everyday beauty, the pursuit of justice tinged with an attraction to danger and a curiosity about criminality, and boldly going where no one has gone before—but embody these concerns very differently from them. This unity within diversity enables the lovers to see how their own values might be realized in ways they probably wouldn't have recognized on their own: moved by Marianne's indignation, Elinor might find it important to stand up to an offense she would otherwise have passively endured; appreciative of Elinor's considered response, Marianne might better understand why her indignation was warranted in the first place. It's even possible that one might not see how one's inchoate jumble of interests and concerns could ground a coherent identity *until* one sees them complemented in a lover, or that one's sense of how to live might become so entwined with one's lover's as to make one lost without them.

More deeply, but more obliquely, there's just something wonderful about someone's picking up on the value you see in your approach to life and your being immediately able to say: "Oh, so it's *not* just me!" Elaborating on Heathcliff's comprehension of her in his own person,

Catherine tells Ellen: "I cannot express it; but surely you and everybody have a notion that there is, or should be, an existence of yours beyond you. What were the use of my creation if I were entirely contained here?" I'm not sure I'm any more able to explain this idea than Catherine was. Still, there seems to be something deeply and intrinsically desirable about communicating to another person who you are and what, as such, is important to you. When you work out your values with a partner, they become more than just the terms of an isolated personal project. Rather, they become intersubjective standards for a way of life that can be lived in common. There's a sense in which values seem more real—more stable and substantial—when they are recognized by another person, and can be examined and assessed from multiple points of view. It doesn't matter to Catherine and Heathcliff that their values be ones that every reasonable person could be expected to share, or even tolerate; bracketing specifically moral considerations, I don't see that it should. But it does, and should, matter to them that their personal values are not *just* personal—that their authority be *inter*subjective.

This point will have to stay at the level of suggestion. Even if it's right, it doesn't yet make it explicit why it matters that the intersubjectivity thus secured have the historical and particular dimension I've claimed to be characteristic of love. To make it more explicit, let's take it from the top. (Conveniently, taking it from the top will also let the discussion do for a conclusion.) If it's true that it's intrinsically desirable that your personal values be more than just personal, this helps explain why it's reasonable to want to be loved selectively, since it would be reasonable to want the affirmation such love would constitute. Such love says, in effect, that you are worthy, at least to *someone*, of special interest and attention *because you identify with the specific values you do*—and if the person in question is someone you love back, you're worthy not just to anyone, but to someone whose judgment really counts. Compare this, again, to Frankfurtian love, which (because arational) does not affirm you in this way; or Vellemanian love, which primarily affirms you,

generically, as a *valuer*; or Kolodnyan love, which affirms you, also generically, as a parent, child, sibling, spouse, or friend.

But the affirmation Catherine and Heathcliff want isn't affirmation from a universal normative perspective. They're too protective of their individuality for that. They want affirmation from a more deeply personal point of view, one that recognizes their values as fundamentally their *own*. Uniquely, someone who loves you as a partner in deep improvisation can provide this more personal kind of affirmation, in loving you for an identity that remains essentially up to you to freely and continually determine. Most of us, fortunately, aren't protective of our individuality in so violent and absolute a sense as Catherine and Heathcliff are. But in wanting to be loved as distinctive individuals, I think we share the same basic concern. In wanting to be loved as *distinctive*, we want our lovers to see, and value us for, aspects of our characters that distinguish us from others. In wanting to be loved as *individuals*, we do not want to be valued merely, as Frankfurt put it, as exemplars of more general types, identifiable and evaluable in abstraction from our particular, concrete, ongoing histories. The improvisational model shows how it is possible to be loved in this way, as persons who are both knowable and endlessly interesting and surprising, with identities that escape determinate categorization but can nevertheless be responded to with fluency and delight.

3

Properly proleptic blame

1 Introduction

In "Internal Reasons and the Obscurity of Blame," Bernard Williams argued that blame involves a kind of "proleptic mechanism."[1] For those to whom, like me, this is all Greek, the *Oxford English Dictionary* reports that *πρόληψις* (*prolepsis*) meant "preconception, especially in Epicurean philosophy, (in rhetoric) anticipation, especially prefigurement, representation of future events, a figure in which objections or arguments are anticipated." I'm going to argue that a central species of blame is proleptic in a very robust sense. It requires that it must sometimes be indeterminate whether you had most reason—internal to your values and commitments—to have performed the action you're blamed for, and anticipates the reasons you may or may not turn out to have upon resolving this indeterminacy. In fact, it is warranted, in part, as part of the process of doing so.

This result is important for two reasons. First, it helps us understand how blame isn't just a bad thing. It plays a vital constructive role in helping people shape their values together. Second, it helps us understand the intimate connection, which many moral philosophers have stressed, between being an appropriate target of blame and traditional questions concerning free will. It's often thought that in order to be appropriately blamed for doing something, you must

1. "Internal Reasons and the Obscurity of Blame," pp. 41-43.

have been able to do otherwise. Recognizing the role of rational indeterminacy in blame lets us capture a more robust sense of this requirement than compatibilists have so far been able to do.

Here is how the argument will go. In the next section, I'll give some reasons to think that there's an important species of blame with two, interdependent features. First, it is constituted by negative emotional responses in roughly the sense P. F. Strawson described in "Freedom and Resentment." But while Strawson focused on guilt and indignation, in addition to resentment, I will focus on anger—and especially on what Marilyn Frye has called "righteous anger."[2] (Following Susan Wolf, who likewise draws on Frye, we can call this kind of blame "angry blame."[3]) I'll then suggest that what makes angry blame distinctive is its expressive, communicative dimension.

In §2, I'll argue that angry blame's communicative dimension makes it inappropriate to address to the sort of agents Williams termed "hard cases"—those with most internal reason to have performed the actions they're blamed for. These agents are in no position to appreciate the concerns expressed in one's blame. If the anger hard cases provoke can be appropriately addressed to anyone, it can only be to third parties; hard cases themselves can only be written off.

In §4, I'll use this result to challenge the view that it's always determinate what, if anything, you have most reason to do. If that were true, then—since we'll already have ruled out hard cases—it would follow that angry blame could *only* be appropriately addressed to agents who had as much or more reason to act better but didn't. In §5, however, I'll argue that this cannot be so.

2. "A Note on Anger."

3. "Blame, Italian Style."

This means we need to reject the assumption that internal reasons are always determinate in order for angry blame to be intelligible. In the final sections, I'll sketch a simple way to do this, and show how the resulting view of blame illuminates its nature and importance.

2 What angry blame is and why it matters

The idea that anger and similar emotions have a special connection to blame is a familiar and natural one. But it is far from obvious what that connection is. For one thing, many perfectly normal instances of blame don't seem to essentially involve negative emotions at all. For instance, if I say you're to blame for getting us lost, I might only mean that you played some especially salient role in *causing* us to get lost—you forgot the map, say. And I might make this judgment without any real opprobrium at all, especially if your lapse was innocent and understandable. If I feel bad about anything, it may just be mild disappointment at the overall state of affairs. (Note, however, that I might nevertheless be perfectly reasonable in expecting an apology from you—and maybe some sort of compensation, too, if getting lost turned out to be really inconvenient.)

It's often thought that what distinguishes this kind of minimal blame for consequences from the kind that essentially involves emotions like anger, resentment, and indignation is that these emotions register attitudes of disrespect or disregard on the part of an offender.[4] Thus, you wouldn't provoke resentment if you got us lost due to an innocent mistake, but you would if you did so out of a malicious desire to amuse yourself at my resulting confusion and stress. However, I don't think things are so simple. There are ways of registering someone's disrespect or disregard for you that are plainly constitutive of at least one important kind of blame, but that do not essentially involve anger or resentment.

4. This may have been Strawson's view; cf. "Freedom and Resentment," p. 14. It is, at any rate, explicitly endorsed by (e.g.) Pamela Hieronymi, in "The Force and Fairness of Blame," p. 135.

T. M. Scanlon's recent theory of blame seems to me to establish this point decisively. For Scanlon, blame consists in the kind of revision of one's attitudes toward other people that it is appropriate to make when those people display attitudes that impair the relationships one has or could have to them. To take Scanlon's example, part of what it is to be friends with someone is to be committed to treating certain considerations as reasons for actions or attitudes toward your friend in ways that do not pertain to others—to judge yourself to have special reasons for being loyal and sympathetic, for instance. But you only have these reasons if the other person is, in fact, really your friend—that is, if the other person is committed to treating you in the same way. If your (putatively) close friend Joe makes fun of you behind your back at a party, he may thereby reveal himself to be no longer your friend, or at least to be less of a friend than you had thought. In such a case, you will not have (all) the reasons for special loyalty and sympathy toward Joe you thought you had. "To revise my intentions and expectations with regard to Joe in this way," Scanlon writes, "is to blame him. I might also resent his behavior, or feel some other moral emotion. But this is not required for blame, in my view—I might just feel sad."[5]

Scanlon's theory thus challenges those who follow Strawson in viewing anger or resentment as essential to (at least some important forms of) blame to clarify just what features of these emotions make them significant, and why. In answering this challenge, R. Jay Wallace and Susan Wolf have both emphasized the way angry blame seems to involve taking a more active, engaged, and vulnerable interest in an offense and the attitudes communicated through it than Scanlonian blame on its own would entail.[6]

I think the most promising approach to understanding this feature of angry blame is in terms of its expressive, communicative dimension. As Margaret Urban Walker puts it, attitudes

5. *Moral Dimensions*, p. 136.

6. See "Dispassionate Opprobrium" and "Blame, Italian Style," respectively.

like anger and resentment "are expressive not only because they reveal something going on in the one who experiences them but also because they are a kind of communicative display that invites a kind of response." Walker persuasively argues that these attitudes are best understood as implicitly *addressed to* others, and as calling out to those to whom they are addressed for "*assurance of protection, defense, or membership* under norms brought in question by the exciting injury or affront."[7] Importantly, resentment and anger, in this sense, can be addressed either to third parties or to offenders themselves. (For what it's worth, "resentment" and "indignation" seem to me mainly to describe attitudes addressed to third parties, while "anger" proper seems mainly addressed at offenders themselves. But of course this is just verbal.)

I won't say much about anger or resentment addressed to third parties, except that this possibility may be worth remembering in what follows, especially for those afraid I might be arguing that you can't appropriately get angry at really terrible people. I'm not. Everything I say below is compatible with the fact that it can make perfect sense to get angry at people like the Koch brothers or the BP oil executives, without presupposing these people to be at all inclined to acknowledge the injustices they advance or commit. I just don't think anger is best understood as addressed to *them*, but rather to third parties. (I'm also willing to be very catholic about who the third parties could be—for all I care, they could be highly general ("the moral community"), merely notional ("the point of view of the universe"), putatively immanent (God), or even, perhaps, oneself alone.) It should be clear, though, that some anger *is* addressed to offenders—and offenders *qua* offenders, not merely *qua* members of the universal moral community.

To get a handle on what it means for anger to be addressed to an offender in this way, it will help to start with a prototypical example. The following one will be important throughout the paper, so I'll present it in detail. It comes from a pivotal scene in *Howards End*. What makes

7. "Resentment and Assurance," pp. 156-157, original emphasis.

this novel interesting in the present context is that its final act can strike one as partly *about* blame. Every major character blames someone at least once, all in different and characteristic ways. By the end the reader is left with a pretty comprehensive catalogue. Helen Schlegel is indignant at her brother-in-law's injustice toward a penniless clerk, and angrily feels betrayed by what she sees as her sister Margaret's complicity in it. Margaret's husband, Henry Wilcox, a sensible man of business, duly revises his attitudes toward Helen in light of what he sees as the impairment to their relationship constituted by her entirely improper anger, combined with her still more improper dalliance with the clerk himself. Of course, Henry blames the clerk too—who for his part is torn apart by remorse—but that blame is old-fashioned retributivism: a man in his position, Henry judges, "must pay heavily for his misconduct, and be thrashed within an inch of his life." Even the third Schlegel sibling, the icily donnish Tibby, faults himself at one point for not being properly reason-responsive: when he betrays his sister's confidence under pressure, he is "deeply vexed, not only for the harm he had done Helen, but for the flaw he had discovered in his own equipment."

Margaret, however, is Forster's heroine, and her blame befits her status. Henry refuses Margaret a small but very important request: to allow Helen to stay with her for her last night in England in his first wife's beloved ancestral cottage. This is not only heartless but hypocritical: Henry had himself been unfaithful to the past Mrs. Wilcox (as it happens, with the woman who went on to marry Helen's clerk). But when Margaret begins to raise that point, his first response is insultingly dismissive:

> "You have not been yourself all day," said Henry, and rose from his seat with face unmoved. Margaret rushed at him and seized both his hands. She was transfigured.
>
> "Not any more of this!" she cried. "You shall see the connection if it kills you, Henry! You have had a mistress—I forgave you. My sister has a lover—you drive her from the house. Do you see the connection? Stupid, hypocritical, cruel—oh, contemptible!—a man who insults his wife when she's alive and cants with her memory when she's dead. A man who ruins a woman for pleasure, and casts her off to ruin other men. And gives bad financial advice, and then says he is not responsible. These men are you. You can't recognize them, because you cannot connect. I've

> had enough of your unweeded kindness. I've spoiled you long enough. All your life you have been spoiled. Mrs. Wilcox spoiled you. No one has ever told you what you are—muddled, criminally muddled. Men like you use repentance as a blind, so don't repent. Only say to yourself: 'What Helen has done, I've done.'"
>
> "The two cases are different," Henry stammered. His real retort was not quite ready. His brain was still in a whirl, and he wanted a little longer.
>
> "In what way different? You have betrayed Mrs. Wilcox, Henry, Helen only herself. You remain in society, Helen can't. You have had only pleasure; she may die. You have the insolence to talk to me of differences, Henry?"
>
> Oh, the uselessness of it! Henry's retort came.
>
> "I perceive you are attempting blackmail. It is scarcely a pretty weapon for a wife to use against her husband. My rule through life has been never to pay the least attention to threats, and I can only repeat what I said before: I do not give you and your sister leave to sleep at Howards End."
>
> Margaret loosed his hands. He went into the house, wiping first one and then the other on his handkerchief. For a little she stood looking at the Six Hills, tombs of warriors, breasts of the spring. Then she passed out into what was now the evening.[8]

It is interesting to compare Margaret's anger with Scanlonian blame. By the end of the exchange, Henry's failure to "connect," as Margaret puts it, has indeed become something Margaret cannot but treat as an impairment to her relationship with him. She lets go of his hands; when she next sees him it is to return his keys and announce her intention to leave him. His refusal has revealed attitudes of his, she has concluded, that make it impossible for her to continue to love him as a husband.

But when Margaret comes to this conclusion, she has *stopped being angry*. She is angry only while it is an open question what Henry will go on to do, and what his initial refusal can mean for them. That her anger is addressed to him as someone who *could still* appreciate the rightness of her case is central to its expressive dimension: she wants him, as Wolf puts it, "to *see* [her] anger and to *feel* [her] pain."[9] When she regards him, finally, as decisively committed to his policy of high-handed indifference, she thereby ceases to regard him as an apt target of her vulnerable, attentive anger. Were Margaret less materially independent and self-assured, her resentment

8. For Emma Thompson's and Anthony Hopkins's enactment of the scene in the novel's Merchant-Ivory adaptation, see `youtube.com/watch?v=07Waj_tL6d4&t=3m`.

9. "Blame, Italian Style," p. 338.

might still call out to others for protection or confirmation; as it stands, there is nothing for it but to withdraw.

3 Why angry blame is inappropriately addressed to hard cases

I now want to argue explicitly that Margaret stops being angry at Henry because she implicitly recognizes that he is (or at least has become) a hard case, and that her anger is as such no longer appropriately addressed to him. But first we need to get clear on the terminology.

Here, I use "hard cases" as shorthand for agents who have most internal reason, all things considered, to perform the actions for which they're blamed. A helpful way to understand what "internal" reasons are, in the present sense, is as reasons "internal" to an agent's subjective values or commitments—they map to what you, personally, value, care about, or find important. (Your "external" reasons, by contrast, map to what you *should* value, care about, or find important, where the stress on "should" is meant to communicate an objective, mind-independent sense.)

Your values, in this sense, are not necessarily identical to your motivations, but they *are* necessarily related to them. How your values call for you to act on a given occasion may differ from how you are motivated to act: you might be motivated to do something that, if better informed, you would recognize as a terrible idea (what you thought was a chocolate chip cookie is actually full of raisins), or *not* motivated to do something that, if you were more mindful and imaginative, you would recognize as marvelously worthwhile. But, as a necessary condition for your values to count in favor of φ-ing, there must be certain non-trivial—if idealized—conditions under which engaging in practical reasoning from your existing motivations could lead you to the conclusion to φ. Needless to say, just what these conditions are is an enormously complicated and voluminously discussed question that I won't begin to answer here. What I'll do instead is use Williams's concept of a "sound deliberative route" as a placeholder for them.

Williams is emphatic that the concept of a sound deliberative route is meant to be very permissive and open-ended. It is meant to cover the wide varieties of ways through which one might come from one's existing motivations to an authoritative conclusion about what one has reason to do. Thus, while sound deliberation presumably includes a correction of errors of matter of fact and reasoning, Williams insists that it's not limited to that. For one thing—and this will be important later—it can also include exercises of imagination and critical reflection. For another, it might also include—as Williams tends to downplay but as other philosophers stress—effective regulation by some subset of one's psychology with authority to represent one's "real" or "deep" self. For now, though, we really don't need anything beyond a vague idea of sound deliberation to go on. All we need to say here is that if you are a hard case, the conclusion that you ought not to have acted badly is not something you can be *reasoned into*—even under ideal conditions, it is not something to be attained through a rational deliberative process, on any plausible specification of the phrase.

Now, the most basic reason to deny that Margaret's angry blame is appropriately addressed to hard cases is simply that it has the distinctive communicative, expressive dimension we saw at the end of the last section. Like other forms of blame, it is a response to the attitudes of disrespect or disregard communicated through the action that provoked it. But Margaret's anger differs from sadness, or even from anger or indignation addressed only to others, precisely in that it involves her thinking or feeling that Henry needs to be shown the wrongness or hurtfulness of his action. (This seems to me to go a long way in explaining how anger is directed *outward* in ways that other emotions are not.) But you can have reason to show something to others—or otherwise communicate something to them—only if you can reasonably hope or expect them to take what you show them into account in deliberation. Yet if Henry lacks a sound deliberative route to appreciating the rightness of Margaret's sense of injury, she cannot reasonably hope or expect

him to properly take it into account. So why should her sense of injury strike her, as it does initially, as something Henry needs to be shown?

This communicative dimension of anger is underscored by what R. Jay Wallace describes as its "element of emotional connection and vulnerability." When "you experience indignation, resentment, or guilt," he writes, "you are not merely left cold by the [offensive] attitudes that form the object of blame, but find that those attitudes engage your interest and attention."[10] Now, Wallace attributes this emotional connection and vulnerability to a special concern for the values or norms against which blamed actions are supposed to transgress (and which he identifies principally with morality). But I think both the phenomenology and importance of anger is better captured by viewing it as reflecting a concern not so much for a set of values, but for the attitudes of the agents to whom it is addressed. Many people who have acted badly toward those they care about *want* those people to get angry at them. If you let someone you love down, an angry response can be positively reassuring. It can mean you're worth the bother, that the person in question really expected more from you and finds your transgression confusing and hurtful. By contrast, when you expect anger and are met instead with detached disapproval, it can be natural to feel shut out or devalued—even, and perhaps especially, when the disapproval is mixed with considerable sympathy and understanding.

But if this is right, it is hard to see how such emotional connection and vulnerability could be appropriate to someone known to be a hard case. The emotions such a person warrants are related to self-protection, rejection, and withdrawal, not openness. Why get wrapped up in a person who harbors disrespect or disregard for you or what you care about? Admittedly, most of us probably *have* addressed our anger to known hard cases at some point in our lives, with all the

10. "Dispassionate Opprobrium," p. 368. I have substituted "offensive" for Wallace's original "immoral," since I see no reason to think that any of these emotions are essentially moralistic.

elements of emotional connection and vulnerability I've associated with it. Nursing some especially nasty insult, the persistent thought of the cruelty expressed can be a nagging pain—something you just keep coming back to, generally with a complicated mix of feelings related to shock, frustration, and disbelief. Reactions like these are very human. But the blame they would constitute seems, at best, to be empty scorekeeping, expressive of a sort of sputtering self-righteousness. I fail to see how it could have any ethical importance that is not as well or better represented by Scanlonian blame alone.

4 Why angry blame is apparently paradoxical

Let's go back to the overall dialectic of the paper. I want to argue, from the existence and intelligibility of angry blame, that it is not always determinate what, if anything, you have most internal reason to do. Now that we have seen why angry blame is indeed inappropriately addressed to hard cases, it should be easy to see how this argument will go. For if angry blame is not appropriately addressed to people whose actions are in accord with their strongest internal reasons, it would be only natural to conclude at this point that angry blame is thus appropriately addressed *only* to agents whose actions are *not* in accord with them. But this looks dubious on its face. For such agents would seem to be acting out of ignorance, confusion, or weakness. Yet one might have thought that the cases in which we get *angriest* at people are those in which their offenses could be characteristic of them. So angry blame seems paradoxical.

It's now time to formulate the argument precisely. Suppose you φ and are angrily blamed for it. I assume that the only relevant possibilities are as follows:[11]

(a) You have most internal reason to have φ-ed.

11. Since it is obvious that changing your mind between acting badly and being blamed for it cannot make blaming you any *more* appropriate, we can assume that your attitudes with respect to *-ing* do not relevantly change over the interval.

(b) You have most internal reason *not* to have φ-ed.

(c) You have just as much internal reason to have φ-ed as not to have φ-ed.

(d) It is indeterminate which of φ-ing or not φ-ing you have most internal reason to have done.

To claim that it is always determinate what, if anything, you have most internal reason to do is to claim that the last of these possibilities never obtains. So assume it doesn't. In the last section, I argued that angry blame is inappropriately addressed to you if (a) is true. It thus follows, given the assumptions, that angry blame could be appropriately addressed to you only if (b) or (c) is true.

I think (c) can be safely set aside as irrelevant. Clearly, it does not pertain to many paradigmatic instances of angry blame, Margaret's blame of Henry included. Henry does not view his choice as a matter of indifference; he sees himself as making a more or less reasoned choice under more or less normal deliberative conditions. If, therefore, it is not plausible that angry blame is appropriate only if (b) obtains, adding (c) will not help. Excluding (d), therefore, would entail that angry blame is appropriately addressed only to agents with most internal reason to have acted as they did. In the next section, I'll argue that this is not so.

5 Why angry blame can't be limited to agents with more reason to act well

It's plausible that angry blame is indeed sometimes warranted in response to faulty or irresolute deliberation. But these can't be the only cases. To begin with, this view of blame would seem to get its emphasis backwards. It suggests that the paradigm cases, when angry blame should be at its strongest and steadiest, are those in which it is clearest that someone's action reflects bad deliberation rather than bad values. But normally the opposite obtains. Strawson suggests that the strength of resentment and indignation "is in general proportioned to what is felt to be the

magnitude of the injury and to the degree to which the agent's will is identified with, or indifferent to, it."[12] Clearly, there's something to this. If, for example, you break a significant (though not vital) promise to a friend because you are clearly not at your best (from being upset or exhausted, say), your offense may be relatively easy to write off. But if you do it deliberately, in a manner that really calls your commitment to the friendship into question, your friend's blame would not, presumably, become correspondingly more tentative. Rather, these are the cases in which angry blame tends to be at its most pressing and insistent.

At this point, other readers of Williams may be inclined to point out that I've left out a central class of cases: those in which blame's proleptic mechanisms come into play. Many people, he notes, may simply "have a motivation to avoid the disapproval of other people—for instance, to avoid their blame." Thus:

> When a motivation of this kind takes a deeper form than merely the desire to avoid hostility, it can be the ethically important disposition to be respected by people whom, in turn, one respects. […] In these circumstances, blame consists of, as it were, a proleptic invocation of a reason to do or not do a certain thing, which applies in virtue of a disposition to have the respect of other people. To blame someone in this way is, roughly, to tell him he had a reason to act otherwise, and in a direct sense this may not have been true. Yet in a way it has now become true, in virtue of his having a disposition to do things that people he respects expect of him, and in virtue of this recognition, which it is hoped that the blame will bring to him, of what those people expect.[13]

Williams reminds us here that there are many perfectly normal reasons for not wanting to be blamed—and, more obviously, for not wanting to provoke the forms of hostile response that many people find naturally expressive of blame. It is typically unpleasant to face someone's anger, and typically much more unpleasant to suffer someone's revenge. Somewhat more subtly, many people simply value avoiding the disappointment or disapproval of others they respect. These are the sort of people Williams describes above. They will have *prima facie* reason to avoid

12. "Freedom and Resentment," p. 21.

13. "Internal Reasons and the Obscurity of Blame," p. 41-42.

doing anything those they respect would blame them, even if those others never express or act on that blame.

The possibility of these desires explain how angry blame can be addressed to agents with most internal reason to act better while still representing the offender's values as potentially defective. For people who value avoiding hostility or embarrassment may have most internal reason, all things considered, to avoid doing things you would blame them for—even if they would not have such reason *but for your anger itself*. So when you address such anger to offenders, you indeed implicitly make what Williams elsewhere calls an "optimistic internal reasons claim"[14]—you hope that, in virtue of your anger, the offenders will appreciate, retrospectively, that they ought not to have acted as they did. And while your anger registers a kind of deliberative mistake—you presuppose that the agent did not properly take into account the fact that you would get angry, or how undesirable your anger would actually be—it does not do so in a way that gets its emphasis backwards. On the contrary, the stronger someone's *independent* reasons for acting badly were, the stronger the *countervailing* reasons constituted by your anger must be if the offender is to be reached.

This point comes out most clearly in the case of anger that does not appeal to an offender's reasons to retain your respect, but rather simply to avoid harm. A great deal of anger, I think, makes a proleptic invocation of *these* reasons in a way that precisely parallels the one Williams describes. (Recall, for instance, Henry's judgment that Helen's clerk should be thrashed for his impudence.) This is the kind of anger associated with a desire to show offenders who's boss, or that nobody messes with you, or that they'll rue the day they crossed you. It aims, just like the other forms of anger I've described, at showing offenders that they in fact had reason not to act badly. But it allows that whether or not this turns out to be the case may ultimately hang

14. "Internal and External Reasons," p. 111.

simply on whether you manage to do things to offenders (or to any other things they care about) that are bad enough to make their offenses not worth it. The stronger their reasons for acting badly would otherwise have been, the worse these things will have to be.

Recall that in the quotation above, Williams suggested that the desire to avoid hostility is less ethically important than the desire to be respected by others one respects. This is presumably so. But the forms of angry blame grounded in each of these desires are much more similar to one another than they are to Margaret's blame, and much less ethically important than it. When you are angry at people, it often does not help when they respond to your expressions of blame by saying things like: "Well, if I knew it *bothered* you so much, I wouldn't have done it." What you want them to see isn't *that* their actions bother you, but *why* they do—or, rather, why you are right in being bothered by them. The reasons you want them to recognize aren't just any reasons not to have acted as they did, but the specific reasons that made their action offensive (their failure to recognize being what justified your blame in the first place). Margaret's anger captures this thought in a way that the kind Williams describes cannot. (To distinguish the two, call the latter *punitive* and—again, following Marilyn Frye—the former *righteous*.)

Thus, while righteous anger is indeed proleptic (or so I will argue), it is not proleptic in the simple sense that punitive anger is. It supposes the agents to whom it is addressed to be in a position to recognize that they ought not to have acted badly *on the same grounds that justified your claim* in the first place, not simply on the grounds that acting badly turned out to be embarrassing or imprudent. Righteous anger, in other words, does not aim at presenting offenders with a new set of considerations, but rather at making vivid existing considerations that they could be expected to appreciate on their own independent merits. For this reason, the features of punitive anger that explain how it can be a response to bad values, rather than (*merely*) bad deliberation, do not apply to righteous anger.

Still, punitive anger has a more positive lesson to teach, which actually gives us *further* reason to think that righteous anger cannot plausibly be limited only to agents with more reason not to have acted offensively. As I explained above, punitive anger represents offenders as agents who really might have had most reason to act as they did *but for your anger itself*. This gives us a start at capturing the sense in which angry blame can represent actions as *threatening*. In general, a threat is something with the potential to harm or undermine something important to you, unless it is somehow met or resolved. Now, as both Strawson and Scanlon have stressed, it is perfectly reasonable for it to be (non-instrumentally) important to you to have a basic level of normative standing in relation to others—that is, that it be important to others to respect the claims you have against them. Actions that occasion angry blame have the potential to undermine this standing. They say, or at least suggest, that offenders did not find your claims against them important enough to respect under the circumstances—and, hence, that they did not find *you*, *qua* holder of these claim, important enough to respect either. Punitive anger recognizes this threat by representing offenders as agents who may lack sufficient independent reason to respect your claims—and aims to resolve it by introducing new considerations that are hoped to tip the scales. The feelings of tension and urgency that anger generally involves, and the experience of it as something to be forcefully expressed outward, thus underscore its distinctive role—as an attitude whose instantiation or expression is necessary to preserve your standing.

The problem is that if *righteous* anger merely registered a deliberative failure, it would have nothing to do in this sense. It could not represents offenses as threats, for it would presuppose your claims to have been important to offenders all along, quite independently of your anger. This strongly reinforces the point about anger's emphasis that began this section: the natural explanation of anger's characteristic urgency and force that punitive anger supports would seem to be unavailable to its righteous cousin. Further, this gives us some positive reason to try to see

how righteous anger could be proleptic as well. Here, appreciating the importance of indeterminacy to blame will help.

6 Internal reasons and indeterminacy

Having ruled out all the options from §4, I've now shown that angry blame is intelligible only if it's not always determinate what you have most internal reason to do. This leaves two questions. First, how could such indeterminacy obtain? Second, how could it help us achieve a positive understanding of what angry blame is like and why it matters?

For a start on the first question, we can take a cue from the way many philosophers of language conceive of indeterminacy in their analyses of vague predicates, like "bald."[15] "Bald," they argue, admits of multiple interpretations, or "sharpenings," each entailing a precise set of necessary and sufficient conditions for when a person is bald. When a claim is true on each sharpening—like "Jean-Luc Picard is bald"—we can say it's determinately true, or "supertrue." When a claim is false on each sharpening—like "James T. Kirk is bald"—we can say it's determinately false, or "superfalse." When it's true on some and false on others—like "Joe Biden is bald"—we can say its truth value is indeterminate.

I want to suggest that a close parallel might hold of the relation between motivations and internal reasons. Remember that for internal reasons theorists, your motivations are not identical to your internal reasons, but rather entail your internal reasons given certain idealizations, which I've called "sound deliberative routes." Here we should keep in mind Williams's insistence on the flexibility of the concept, and the corresponding point that there are all sorts of things we might want to build into it, above and beyond correction of errors of fact and reasoning—exercises of imagination, critical reflection, effective regulation by attitudes with authority to represent the

15. Cf. e.g. Kit Fine, "Vagueness, Truth, and Logic."

agent's real self, whatever. This is, again, no time to get into details. The point I want to make now is that whatever the details turn out to be, it's perfectly conceivable that they could be such as to support multiple, mutually incompatible idealizations of an agent's motivations at a given time, in precisely the same way that vague predicates might admit of multiple, mutually incompatible sharpenings. When this happens, what the agent has most internal reason to do will be indeterminate.

Though it hasn't gotten much attention from his many critics and commentators, Williams himself stresses the possibility and importance of rational indeterminacy, in both "Internal and Reasons and the Obscurity of Blame" and its more famous prequel:

> If someone is good at thinking about what to do, he or she needs not just knowledge and experience and intelligence, but imagination; and it is impossible that it should be fully determinate what imagination might contribute to deliberation. This is one reason why it may be indeterminate what exactly an agent has reason to do.[16]

Williams's appeal to imagination gives us a simple, concrete way of understanding the indeterminacy angry blame requires. As he explains in "Internal and External Reasons," imaginative deliberation can itself influence the motivations from which one deliberates and chooses:

> [An agent] may think he has reason to promote some development because he has not exercised his imagination enough about what it would be like if it came about. In his unaided deliberative reason, or encouraged by the persuasions of others, he may come to have some more concrete sense of what would be involved, and lose his desire for it, just as, positively, the imagination can create new possibilities and new desires. (These are important possibilities for politics as well as for individual action.) […] We should not, then, think of S [that is, the agent's subjective motivational set] as statically given.[17]

Interestingly, David Lewis considers the same possibility in "Dispositional Theories of Value," but—in telling contrast to Williams—claims that in "ideal" conditions of imaginative

16. "Internal Reasons and the Obscurity of Blame," p. 38.

17. "Internal and External Reasons," pp. 104-105.

acquaintance, no such change in the agent's motivations will occur.[18] Whether or not such conditions are ideal, however, they obtain all the time. Suppose, for instance, you are deciding whether to spend an afternoon doing some much-needed work on your dissertation or helping your friend move. If you consider the dissertation first, you may imagine it in ways that make the considerations in favor of helping your friend more salient, weigh these considerations more heavily going forward, and so settle on helping your friend. Yet if you started with that option you would go *mutatis mutandis* for the dissertation. What do you have more internal reason to do? (The one you are disposed, *right now*, to consider last?)

Note that there is no reason to think that anything about your psychology *per se*—much less whatever portion of it gets to count as your "subjective motivational set"—suffices to determine which deliberative route you will take. Presumably, which deliberative route you will take is sensitive to all sorts of totally incidental facts about your physiology and environment. However decisively you settle on helping the friend, it may yet be the case that had only your eyes not settled on your photocopy of "Truth, Invention, and the Meaning of Life," with those nice passages about digging a ditch with a man whom one likes, you might have no less decisively settled on dissertating. So there is no single, univocal answer to the question of what you have most reason to do, because there is no single, univocal answer to the question of what conclusion you will reach from your existing motivations by a sound deliberative route. *Relative to your motivations prior to deliberation*, it is neither the case that you have most reason to help your friend, nor that you have most reason to dissertate, nor that you have equal reason to do each. What you have most reason to do is indeterminate.

18. "Dispositional Theories of Value," pp. 121-126.

7 Properly proleptic blame

To apply this result to blame, let's go back to *Howards End*. At the beginning of her exchange with Henry, Margaret implicitly saw him as someone whose relevant internal reasons either were, or may well have been, indeterminate—that is, as "muddled, criminally muddled," to use her now-strikingly apt phrase. Margaret does not love him foolishly—he is, in many ways, a thoughtful and humane person—but he also has what Forster calls his "fortress"—a perspective defined roughly in terms of values related to dominance, self-sufficiency, and, at best, a vaguely patriarchal sense of honor. Depending on how these sides of his personality came into play, he may well have had sound deliberative routes from his initial motivations both to the conclusion to allow Margaret and Helen to stay at Howards End and to the conclusion he actually reached. Had things only been different, in ways incidental to his motivations proper, Margaret may well have reached him.

By the end of the exchange, however, it is no longer plausible that this is so. Henry has firmly resolved to treat his wife's request as an attempt at blackmail, and he is emphatically not the sort of person who would willingly abandon such a resolution having made it. His stance has hardened and his attitudes have shifted. If, in letting go of his hands, Margaret implicitly concluded that there is no longer a sound deliberative route from these attitudes to the recognition of her demand as just, she would have been eminently reasonable in doing so.

Between these two points, Henry is engaged in an ongoing deliberative process. He is trying to formulate a principled justification for his refusal, in a form he can articulate to Margaret. In trying to decisively settle on this justification, and thereby commit himself motivationally to it, Henry is literally in the process of *making determinate* where he stands with respect to Margaret's request. Margaret's blame implicitly registers this fact in aiming to influence the process. It reflects the thought that if Margaret could just make vivid enough to

Henry how she sees his refusal, he might—by properly taking what her anger shows him into account—turn out to have most reason not to have made it.

This allows us to explain how righteous anger can be proleptic in a way that neatly parallels its more primitive cousin. Like punitive anger, it is a kind of optimistic internal reasons claim. It addresses offenders—optimistically—as agents who may indeed have most internal reason to act better, but also as ones who might not have had such reason *but for your anger itself*. The "but for your anger itself" clause, however, means something very different for righteous anger than it did for the punitive kind. The latter, as we've seen, allows that the agent may (determinately) lack independent reason to act well, and so is proleptic only in the sense of introducing considerations of its own (i.e., that the offense will turn out to be embarrassing or imprudent) that are hoped to tip the scales. Righteous anger, by contrast, does not aim at introducing new considerations, but rather at playing a (causal) role in the agent's process of resolving the indeterminate status of existing ones by making their force clearer and more vivid to the agent.

In fact, Williams put his finger on this kind of proleptic mechanism, too. He writes:

> Our thought may rather be this: if [the offender] were to deliberate again and take into consideration all the reasons that might now come more vividly before him, we hope that he would come to a different conclusion; and it is important that the reasons that might now come more vividly before him include this very blame and the concerns expressed in it. This kind of thought helps us to understand a sense in which focussed blame asks for acknowledgment.
>
> A rather similar structure can apply to advice…[f]or even when we are advising in the 'if I were you' mode, our claim that the agent has most reason to φ does not necessarily mean that simply given his S as it is, it already determines that φ-ing has priority over anything else. We are saying that the conclusion to φ, rather than to do something else, can be reached from his S by a sound deliberative route, and that is something that involves such things as the exercise of his imagination and the effective direction of his attention. But among the things that will affect his imagination and his attention, we hope, is our advice itself and how it represents things.[19]

There is one important difference between righteous anger and advice, however, that Williams does not discuss. The possibility that it might be rejected (as Margaret's of course was) has a very

19. "Internal Reasons and the Obscurity of Blame," p. 42.

different significance for the former than the latter. In angrily blaming Henry, Margaret regards him as someone who *might* be reached, but also as someone who might not. The force of her anger registers what is for her the very real possibility that Henry's offense could turn out to be characteristic of him. Thus, Margaret's anger is two-faced in that it regards Henry as occupying a kind of unstable, liminal position, halfway between that of a hard case and someone who merely made a deliberative mistake. Perhaps uniquely, it represents its object as warranting both rational appeal—that is, as someone potentially still in a position to acknowledge the rightness of your demand—*and* defensiveness or hostility—that is, as someone potentially beyond the reach of reason, and who can only be opposed. Margaret wants, very badly, for Henry to turn out to be someone whose action appears, retrospectively, as a deliberative mistake. As such she still presents *reasons* to him, albeit as forcefully and vividly as she can. But the fact that he might resolve his indeterminacy the other way gives her anger its edge. He could turn out to be someone to whom she cannot reach out, but whom she must rather forcefully reject, and it is important that her anger be something that could be part of doing either one of these things. To this end it says, in effect: "*This* is what your action means to me. Will you acknowledge it? If so, good. If not, you've been warned. The ball's in your court."

8 Conclusion

I want to finish by flagging a pair of what strike me as noteworthy implications of my view of blame. First, viewing righteously angry blame as proleptic appealingly highlights its constructive role. It does not just provide a way for people to enforce against each other the terms on which it is important to them to interact, but enables them to jointly shape and refine just what those terms are. That is, suppose it is true—as seems at least *prima facie* plausible—that many of the details concerning how it is important to people to live, and the terms on which it is important to

interact with those connected to them in various ways, are indeed indeterminate at any given point. When you get angry at someone for violating the terms of a relationship as you understand them, there may be no fact of the matter whether the person you are angry at valued your relationship on *quite* those terms, or saw the action as inappropriate on quite the same grounds that you did, or to quite the same extent. Your blame calls on the person to fill in those details. In this, it provides a way of achieving common ground that neither simply reminds people of shared values or commitments they had all along, nor goads them into adhering to norms whose authority they could not appreciate through deliberating from their existing motivations.

As a bonus, this feature of angry blame helps explain how it can make sense to get angry at people whom you're sure really did have most reason to have acted better. These cases can, and often do, at least raise the question whether the person you're angry appreciates the reasons in favor of acting better in the same way that you do. Anger between close family members, friends, or co-workers offers a lot of examples here. Suppose you commit some venial negligence: you over-water, and kill, a plant you promised to tend for a friend. Clearly, your action does not raise any question whether you in fact had most reason to take good care of the plant—of course you did. But it can raise questions about the specific respects in which promises of that kind—or plants, for that matter—are important to you in that context. These are the kind of questions that a great deal of angry blame helps settle, by communicating just what such actions mean to the people who suffer them, and calling on offenders to affirm or deny the concerns their blame expresses.

Second, and more speculatively, the indeterminacy involved in blame suggests a promising analysis of its evidently deep but obscure connection to freedom. Incompatibilists have long insisted that blame requires a power for opposites—that you can appropriately be blamed for acting badly only if you could have done otherwise. Compatibilists have long responded by

offering analyses of the power for opposites that show it to be compatible with determinism—surely, they suggest, it simply concerns some perfectly reasonable requirement of self-control (such that one could have done otherwise had one wanted to), or of some adequate opportunity to avoid (such that one could reasonably be expected to have done otherwise, as one couldn't, for instance, if coerced). And incompatibilists have long replied, in turn, that these analyses are missing the point: the requisite power for opposites must be more robust than that. What blame requires, they standardly insist, is that it must have been possible for the agent to do otherwise—and to do so intentionally and attributably—given all the facts about the world and the laws of nature up to the bad action or the decision to perform it.

My view of blame does not entail anything quite so robust. But it does entail something strikingly close. It holds that for angry blame to be fully intelligible, there must be cases in which it would have been possible for the agent to do otherwise—again, intentionally and attributably—given all the facts about the agent's *motivations* up to the bad action (and, for that matter, for some time afterwards). I don't think this is a coincidence. I rather think that one of the major reasons the phenomenology of blame can make incompatibilism seem intuitive is that angry blame really does represent it as having been open to the agent, in a very deep sense, to have acted better. This is what gives blame its drama, its sense of uncertainty and tension. I've offered a way to capture this drama in terms of blame's relation to emotions, relationships, and practical rationality—"to recover it from the facts as we know them," as Strawson put it, "without recourse to the obscure and panicky metaphysics of libertarianism."[20]

20. "Freedom and Resentment," pp. 23 and 25.

4

Two concepts of self-creation

1 The elusive appeal of self-creation

In discussions of free will in analytic philosophy, the following dialectic seems to me common. The incompatibilist produces an argument, like Peter van Inwagen's Consequence Argument or Derk Pereboom's manipulation arguments, that purports to show that determinism is intuitively incompatible with some *prima facie* significant feature of agents—such as the ability of agents to do other than what they do, in van Inwagen's case, or the ability to be morally responsible for what they do, in Pereboom's—where the feature in question is considered prior to substantive analysis.[1] The compatibilist responds by actually producing a substantive analysis of the feature in question—as (for instance) David Lewis and Hilary Bok have done with respect to the ability to do otherwise, and John Martin Fischer, R. Jay Wallace, and T. M. Scanlon have done with respect to moral responsibility—that either explains away the incompatibilist's intuition or challenges its credibility.[2] Retrenching, the incompatibilist then either makes the negative argument more elaborate, or insists, through a basic appeal to intuition, that the compatibilist's substantive analysis is superficial.

I will not engage this dialectic in this chapter. Even if (as I in fact believe) existing compatibilist accounts of the ability to do otherwise and of moral responsibility are at least

1. See, respectively, "An Argument for Incompatibilism" and *Living Without Free Will*.

2. See, respectively, "Are We Free to Break the Laws?" and "Freedom and Practical Reason," concerning to the ability to do otherwise, and *Responsibility and Control: A Theory of Moral Responsibility* (coauthored with Mark Ravizza), *Responsibility and the Moral Sentiments*, and "The Significance of Choice," concerning moral responsibility.

approximately correct, it seems to me that they have left unaddressed what may be the deepest, most profound, and most elusive source of the sense that determinism threatens something vital about human agency. Thus, Robert Nozick writes:

> Philosophers often treat the topic of free will as a problem about punishment and responsibility: how can we punish someone or hold him responsible for an action if his doing it was causally determined, eventually by factors originating before his birth, and hence outside his control? However, my interest in the question of free will does not stem from wanting to be able legitimately to punish others, to hold them responsible, or even to be held responsible myself.
>
> Without free will, we seem diminished, merely the playthings of external forces. How, then, can we maintain an exalted view of ourselves? Determinism seems to undercut human dignity, it seems to undermine our value.[3]

For Nozick, like many others, what determinism seems above all to threaten is the valuable status of being the "originators of our acts" and of the value we realize in acting. If it were not fundamentally up to us what to do, and ultimately who to be, our importance as individuals would somehow be undermined. "Free will has been traditionally conceived as a kind of creativity," Robert Kane writes, "akin to artistic creativity, but in which the work of art created is one's own self. As ultimate creators of some of our own ends and purposes, we are the designers of our own lives, self-governing, self-legislating—masters, to some degree, of our own moral destinies." It is plausible, Kane argues, "that underived origination or sole authorship is necessary for a number of other things that humans generally desire and are worth wanting." These things include, in part, creativity, autonomy, desert, moral responsibility "in an ultimate sense," dignity or self-worth, "a true sense of individuality or uniqueness as a person," "life-hopes regarding an open future," and love and friendship.[4]

But the trouble with self-creation is that it has stubbornly resisted analysis. Attempts to spell out the idea of self-creation can result, all too easily, in something that is either unintelligible or inadequate to its pretheoretic importance. Compatibilists and hard determinists alike have

---

3. *Philosophical Explanations*, p. 291.

4. *The Significance of Free Will*, pp. 81-89.

long argued that this result is inevitable. "The logical goal of [the] ambitions" intrinsic to the intuitive idea of autonomy, Thomas Nagel argues, "is incoherent, for to be really free we would have to act from a standpoint completely outside ourselves, choosing everything about ourselves, including all our principles of choice—creating ourselves from nothing, so to speak. This is self-contradictory: in order to do anything we must already be something."[5] While hard determinists stolidly conclude that we are indeed condemned to want something impossible, compatibilists are somewhat more optimistic. They argue that the concern for self-creation was misguided from the start, and we should just get over it. Harry Frankfurt's discussion of the matter is paradigmatic, and warrants extended quotation:

> A person's will is real only if its character is not absolutely up to him. It cannot be unconditionally within his power to determine what his will is to be, as it is within the unconstrained power of an author of fiction to render determinate—in whatever way he likes—the volitional characteristics of the people in his stories. […] Remember Hotspur's reply when Owen Glendower boasted, "I can call spirits from the vasty deep." He said: "Why, so can I, or so can any man; but will they come when you do call for them?" The same goes for us. We do not control, by our voluntary command, the spirits within our own vasty deeps. We cannot have, simply for the asking, whatever will we want.
>
> We are not fictitious characters, who have sovereign authors; nor are we gods, who can be authors of more than fiction. Therefore, we cannot be authors of ourselves. Reducing our own volitional indeterminacy, and becoming truly wholehearted, is not a matter of telling stories about our lives. Nor, unless we wish to be as foolish as Owen Glendower, can we propose to shape our wills by stipulating peremptorily at some moment that we are now no longer divided but have become solidly resolute. We can only be what nature and life make us, and that is not so readily up to us.
>
> This may appear to conflict with the notion that our wills are ultimately free. A natural and useful way of understanding it is that a person's will is free to the extent that he has whatever will he wants. Now if this means that his will is free only if it under his entirely unmediated voluntary control, then a free will can have no genuine reality; for reality entails resistance to such control. Must we, then, regard our wills either as unfree or as unreal?
>
> This dilemma can be avoided if we construe the freedom of someone's will as requiring, not that he control or originate what he wills, but that he be wholehearted in it. If there is no division within a person's will, it follows that the will he has is the will he wants. His wholeheartedness means exactly that there is in him no endogenous desire to be volitionally different than he is. Although he may be unable to create in himself a will other than the one he has, his will is free at least in the sense that he himself does not oppose or impede it.[6]

5. *The View from Nowhere*, p. 118.

6. "The Faintest Passion," p. 101-102. While Frankfurt argues that the requisite attitudes of higher-order endorsement are a species of desire, Michael Bratman has developed an important alternative hierarchical model, in which the attitudes in question are "self-governing policies," which for Bratman are species of plans, rather than desires. (See e.g. his *Structures of Agency*, esp. chs. 2, 4, 7, and 8.)

Frankfurt is not being dismissive here. A unifying theme of his writing on free will, very much in play in this passage, is that "classical" compatibilists like Ayer, who argue that free will worth the name could only amount to the freedom to do what you want, really are missing something vital. There is something deeply right about the thought that freedom is not just free *exercise* of the will, as the classical compatibilists thought, but freedom to have the will you want. The mistake, Frankfurt argued, is in the thought that in order to have the will you want, you must have *created* your will. Instead, your will is free if it is, in fact, the will you endorse at a higher level, and if you are unambivalent about this endorsement (in the sense that there is nothing you would see as a reason to oppose it).

However, it has often been argued that "hierarchical" theories like Frankfurt's *still* miss something vital. For instance, if the requisite higher-order endorsement comes about in the wrong way, or involves the wrong attitudes, it can reasonably seem to an outsider that its presence is insufficient to really mark out the agent as free. As Susan Wolf has argued, for instance, people like dictators and reactionary fanatics can plausibly be described as *trapped* in worldviews to which they are nevertheless reflectively and wholeheartedly devoted, in a way that adequate theories of free agency should be able to account for. Compatibilists have developed this point in two directions, both ways of arguing, as Frankfurt did, that what might have *seemed* like a reason to insist that you must create yourself to truly be free actually isn't one. John Martin Fischer, for instance, has argued that hierarchical conditions need to be supplemented with a further historical condition: while it is not necessary to have exercised ultimate control over the process through which you come to have the will you do, it must be a process you can "take responsibility" for, as you could not if your will results from manipulation or other morally

suspect means.[7] Wolf herself holds that freedom requires a capacity for *self-correction*, rather than self-creation, constituted by a standing commitment to revise your higher-order motivations through the exercise of a "minimally sufficient ability cognitively and normatively to recognize and appreciate the world for what it is."[8]

The implications of such a self-correction condition are not limited to the conditions under which someone may be recognized as free by a third party. They also extend, importantly, to what agents presuppose about their own practical reasoning when they regard themselves as acting freely: namely, that they aim to track what is in some way objectively valuable. This in itself marks a striking change from Frankfurt: to have free will is no longer just to have the will *you* want, but for your will to be responsive to independent standards for what it should be. Philip Pettit and Michael Smith usefully name this kind of freedom *orthonomy*, as opposed to *autonomy*—right rule, as opposed to self-rule—and argue on its basis that freedom of the will is essentially similar to freedom of thought. To think and will freely is for your beliefs and desires to be determined by objective standards of theoretical and practical rationality, rather than by arbitrary external causes.[9] This is hardly a new thought: versions of it have been defended by Plato, Kant, and probably above all Spinoza.[10] Nagel himself proposes one, as a sort of "next

7. See John Martin Fischer and Mark Ravizza, Responsibility and Control: A Theory of Moral Responsibility, chs. 7 and 8.

8. "Sanity and the Metaphysics of Responsibility," p. 381. Many philosophers follow Wolf in the view that some threshold of normative competence is necessary for the freedom, though it remains controversial what the threshold should be and what kind of considerations bear on its position. One especially influential approach, which follows P. F. Strawson in viewing freedom primarily as a condition for moral responsibility, argues for identifying the threshold with the minimal degree of normative competence necessary for someone to be fairly held to moral demands. See e.g. R. Jay Wallace, *Responsibility and the Moral Sentiments*, T. M. Scanlon, "The Significance of Choice" and *Moral Dimensions*, Stephen Darwall, *The Second-Person Standpoint*, Gary Watson, "Responsibility and the Limits of Evil," and Philip Pettit and Michael Smith, "Freedom in Belief and Desire."

9. See their "Freedom in Belief and Desire." It is probably fair to say that the orthonomy view is the default position in philosophical discussions of agency. See e.g. Gary Watson, "Free Agency," Angela Smith, "Responsibility for Attitudes; Activity and Passivity in Mental Life," and Pamela Hieronymi, "Controlling Attitudes."

10. In "The Makropulos Case: Reflections on the Tedium of Immortality," Bernard Williams touches on the

best thing" to origination. Even though we cannot create ourselves, he offers, we can at least aspire to transcend our contingent preferences and predispositions by choosing according to objective evaluative standards (which inevitably turn out to include, if not be identical to, the standards of morality). While you may still be determined by these preferences and predispositions, you need not feel *trapped* by them, since you may at least hope to be acting on grounds that would rationally command endorsement from anyone, regardless of whether they had the preferences or not. Or at least, more realistically, you may hope to be acting on grounds that would not be *rejected*. And even in "these kinds of cases," Nagel notes, "I do not feel trapped or impotent when I consider my situation objectively, because I do not aspire to more control than I have if my choice is dictated by my immediate inclinations. I am content with the freedom of a cat choosing which armchair to curl up in. External assessment can add nothing to this, nor does it detract."[11]

I do think the hierarchical view misses something vital, but I also think the orthonomy view fails to appreciate the extent of the problem. The considerable differences between these views belie a shared allegiance to a very general picture of freedom as a kind of *receptivity*. On this picture, freedom is a matter of bringing the will into conformity with rational standards independent of it. The two views differ only with respect to the sources of these standards. On

---

Spinozist formulation of the view, as expounded by Stuart Hampshire, before voicing dissatisfaction with it on intriguing and characteristic grounds: "I am not sure that the Spinozist consideration which Hampshire advances even gives a very satisfactory sense to the *activity* of the mind. It leaves out … the driving power which is needed to sustain one even in the most narrowly rational thought. It is still further remote from any notion of creativity, since that, even within a theoretical context, and certainly in an artistic one, precisely implies the origination of ideas which are not fully predictable in terms of the content of existing ones. But even if it could yield one sense for 'activity', it would still offer very little, despite Spinoza's heroic defense of the notion, for *freedom*. Or—to put it another way—even if it offered something for freedom of the intellect, it offers nothing for freedom of the individual. For when freedom is initially understood as the absence of 'outside' determination, and in particular understood in those terms as an unquestionable *value*, my freedom is reasonably not taken to include freedom from my past, my character, and my desires. To suppose that those are, in the relevant sense, 'outside' determinations, is merely to beg the vital question about the boundaries of the self, and not to prove from premises acceptable to any clear-headed man who desires freedom that the boundaries of the self should be drawn around the intellect." (p. 97)

11. *The View from Nowhere*, p. 131.

hierarchical views, the sources are psychological states necessarily resistant to the agent's direct voluntary control; on orthonomy views, they are objective normative facts. Both views unapologetically affirm what William James derided as a "a *soft* determinism which abhors harsh words, and repudiating fatality, necessity, and even predetermination, says that its real name is freedom; for freedom is only necessity understood, and bondage to the highest is identical to true freedom."

In this, both hierarchical and orthonomy views dismiss as misguided, if not unintelligible, what might have seemed to be one of the most basic and compelling incompatibilist intuitions: that freedom of the individual essentially requires that it be ultimately *up to you* to determine who you really, fundamentally are. On hierarchical views, who you are is no more up to you to determine than it was up to Owen Glendower whether the spirits from the vasty deep come when he calls them. On orthonomy views, who you are is up to you to determine to precisely the same extent that what to believe is. And while this is not nothing—freedom of thought is a more than legitimate object of aspiration—I find it incredible that anyone could seriously suppose epistemic freedom to be the *same* kind of freedom yearned for in what Kane describes as the desire to be self-governing, self-legislating designers of our own lives. (If the idea of being the designer of my own *epistemic* life makes sense at all, it's certainly not anything I want. I am more than happy to cede that prerogative to the world.) It may turn out, at the end of the day, that Kane's desire is not actually for anything worth wanting, but it can't be so obviously misplaced.

I will argue that the only way to get a grip on the nature and importance of Kane's sort of self-creation is to take seriously the idea that free agents create themselves in a radical sense. When compatibilists argue that what might have looked like a reason to value self-creation is actually just a reason to value something else, like reflective endorsement or orthonomy, they

really are to be faulted for changing the subject. But I will also argue that what is indeed a reason to value self-creation may not yet be a reason to value something *incompatible with determinism*.

I'll begin by focusing on the basic idea that free agents are the originators of their actions, in roughly the sense Thomas Nagel describes as the intuitive idea of autonomy. Nagel presents powerful reasons to think that if autonomy indeed requires origination, it must also require a radical form of self-creation. For autonomous actions must not only originate with you, but also be *attributable* to you, in the sense of being intelligible as things *you* do, rather than ones you randomly find yourself doing. But an action can both originate with you and be attributable to you only if you somehow *create* the "self" to which it is to be attributed through making it. As we've seen, Nagel thinks this is impossible. But I will argue that it is only impossible given a certain substantive, broadly causal conception of the self as the source of agency. But this is not the only conception according to which we might interpret the intuitive idea of autonomy. By reinterpreting the idea in normative terms, we can see how it might require a form of self-creation that is both intelligible and intelligibly important—and also compatible with determinism.

2 The intuitive idea of autonomy

Many of the ideas investigated in philosophy are straightforwardly derivable from pretheoretically compelling commonplaces, or from the presuppositions of prominent attitudes and practices. That a vital form of autonomy essentially involves self-creation or something like it is not one of them. This is much of what makes it so difficult to get a grip on it. The closest thing to a starting point, it seems to me, is an inchoate sense that it is somehow important that what you do is somehow ultimately up to you—that your choices about what to do and how to live are *yours*, and *only* yours, to make. Call this vital capacity to make choices that are yours and only

yours *autonomy*. (Given the prevalence of the word in many domains of ethical and political discourse, it should probably be stressed just in case that I do not mean to give a general theory of autonomy. I take the autonomy putatively threatened by determinism to be one of may species of it, and not necessarily the most basic or important one.)

Because disputants in philosophical discussions of free will are so often accused of changing the subject, I think it is important to begin by recognizing that autonomy really is awfully vague. What we know starting out is that we sometimes seem to be capable of making choices that are in some robust sense ours and only ours to make, that this capacity seems somehow deeply important, and that something about this capacity can easily seem to be threatened by determinism. If we built anything more into autonomy than this, we risk unselfconsciously distorting it and making our job impossible before we start. All this is a consequence of the fundamental feature of the free will problem: namely, that it is a *problem*, not just an idle intellectual puzzle. It keeps people up at night, throws profound thinkers into years of distress, and drives anxious adolescents into philosophy classrooms. It arises because determinism seems to threaten something vital, and no philosophical analysis of autonomy can constitute an adequate response to the free will problem unless the features of agency putatively constitutive of it really are recognizable as vital, and vital in the right way. But what counts as "vital in the right way" is, to put it mildly, anything but obvious.

This brings us to Nagel. Nagel's discussion of autonomy offers a good point of departure because Nagel is characteristically sensitive to the subtlety and importance of the issue, and the dangers of forcing it into an artificially narrow mold. When he describes "what [he] take[s] to be our ordinary conception of autonomy," he does so in explicit acknowledgment of the contestability of his interpretation. I think he comes as close as anyone to getting at the heart the matter. But I also think that even he specifies too much too soon. Nagel writes:

> [Our ordinary conception of autonomy] presents itself initially as the belief that antecedent circumstances, including the condition of the agent, leave some of the things we do undetermined: they are determined only by our choices, which are motivationally explicable but not themselves causally determined. Although many of the external and internal conditions of choice are inevitably fixed by the world and not under my control, some range of open possibilities is generally presented to me on an occasion for action—and when by acting I make one of those possibilities actual, the final explanation of this (once the background which defines the possibilities has been taken into account) is given by the intentional explanation of my action, which is comprehensible only through my point of view. My reason for doing it is the *whole* reason why it happened, and no further explanation is either necessary or possible.[12]

What I think is right in Nagel's description is that it captures, relatively simply and directly, the two essential elements of autonomy. I'll call these *origination* and *attributability*: a choice or action is autonomous (that is, it manifests autonomy) only if it both *originates with* you and *is attributable to* you. As a first pass, a choice originates with you if all of the facts about your circumstances and yourself—antecedent to and independent of the choice itself—leave it open to you whether or not to make it; an action originates with you just in case the choice to perform it does. When you are the originator of your actions, then, what you do depends on what you choose to do, and the antecedent conditions do not determine the choice you will make. Similarly, when we try to *explain* an action that originates with you, that you here and now *chose* to do it for the reasons you did constitutes an ineliminable part of the explanation. (This is the "up to *only* you" part.)

If autonomy only required origination, it might be contingently impossible, but it would not be conceptually problematic. Notoriously, what makes autonomy conceptually problematic is that origination in this basic sense is clearly not enough for it. As Nozick puts it, "if an uncaused action is a random happening, then this no more comports with human value than does determinism. Random acts and caused acts alike seem to leave us not as the valuable originators of action but as an arena, a place where things happen, whether through earlier causes or spontaneously."[13] That is, there must be a difference between what you freely do and what you

12. *The View from Nowhere*, pp. 114-115.

13. *Philosophical Explanations*, p. 292.

randomly happen to do, but origination in the basic sense above does not mark this. Call this difference, whatever it turns out to be, "attributability." Nagel takes attributability to be a matter of the kind of explanation of which autonomous actions admit. This is plausible, because an action must be explicable in one way or another for it to be intelligible why you are doing it, and to the extent that it is unintelligible to you why you are doing something, what you doing does not seem to be under your control. (*Mutatis mutandis* for choices.) This point is especially vivid in the context of one's experience of one's own activity. If I unaccountably got up from my chair and began walking down the hall, I would not experience this as a meaningful exercise of autonomy, regardless of whether my doing so was underdetermined by antecedent conditions. It would be as though a bizarre force had taken control of me. The addition of a brute desire to do so ("must...walk…thusly…") would not help. (This is the "up to *you*" part.)

All of these features of origination and attributability, I will argue, are indeed essential to the vital form of autonomy that determinism seems to threaten, and that hierarchical and orthonomy views neglect. But Nagel's description goes a step further, and here things get complicated. Nagel distinguishes between two kinds of explanation of action: causal explanations and what he calls "intentional" explanations, which are "in terms of justifying reasons and purposes."[14] On his view, autonomous actions necessarily admit of "intentional" explanations but not causal explanations, or at least not complete causal explanations. I do not think this is quite right—at any rate, not as an *initial* formulation of autonomy—although I do think there is an element of truth to it, and that Nagel's distinction between kinds of explanation is absolutely critical. For, as I will now explain, we can distinguish two senses in which an action or attitude

14. *The View from Nowhere*, p. 115. "Intentional" is in scare-quotes because Nagel's terminology is infelicitous. Plausibly, an action can be explicable in the sense required for it to count as intentional without necessarily being explicable in the sense required for it to count as a meaningful exercise of autonomy. (In fact, Nagel's own argument depends on this.) In later sections, I will use the phrase "justifying explanation" instead.

can be understood as attributable to an agent, each prominent in theorizing about agency and freedom, and each requiring the possibility of different kinds of explanations as constitutive of attributability. These senses of attributability are best understood as part of larger conceptions of the self as a source of agency. One conception is (broadly speaking) *causal*; it holds that something is attributable to you if it is caused by you in the right way. The other is *normative*; it holds that something is attributable to you if it is embodies (what we can call) your normative self-conception.

A fruitful way to approach the difference between these conceptions of attributability is in terms of the different kinds of questions about relationships between agents and their actions they purport to answer. Each kind of attributability is important, but each kind is important in very different contexts from the other, and for very different reasons. What's striking, however, is that questions about *each* kind of attributability can introduce questions that motivate a corresponding kind of origination. To take the parallel even further, each of these kinds of origination are subject to structurally similar challenges, each ultimately pointing toward self-creation. Nevertheless, the nature and potential significance of the self-creation in question differ completely between the cases. I think Nagel's argument against the possibility of autonomy actually runs these strands together, and that disentangling them will make it easier to see what autonomy could be and why it might be so compelling. So that is what I will do in the next two sections, starting with the causal case, which is the simpler of the two.

3 The causal conception

Billy throws a rock; it hits the window; the window breaks. Interested in whether and to what extent to hold him accountable for this event, we might naturally want to know the degree of control he had over its occurrence, and whether it resulted in states or dispositions stable or

ingrained enough for us to reasonably expect that Billy might do similar things in different situations. The answers to these questions depend on the causal explanation of how the breaking of the window came about, and specifically on Billy's position in it. We can ask: was the event of the window's breaking an intended or foreseeable consequence of something Billy did intentionally? (Maybe he was throwing the rock in a completely different direction, as part of a game, and was redirected towards the window in the air through Suzy's masterfully-placed throw.) If it was, we can ask: did Billy have a normal degree of control over what he did? (Maybe Suzy's neuroscientist father, Black, put a device in his brain.) And if he did, we can still ask: did he have a normal degree of control over his *motivations* for doing what he did? (Maybe Black had been forcing him to consume a lot of anarchist propaganda, in a kind of reverse *Clockwork Orange* scenario.)

The first of these questions concerns Billy's relation to an event; the second, to an action (as opposed to an event only contingently related to an action, as in the prior case); the third, to a set of attitudes. It is possible, and in some contexts proper, to approach all of them in the same way, as admitting of the same kind of answer. We can say that the event of the window's breaking is relevantly attributable to Billy to the extent that he had some appropriate degree of control over it. Here we can assume that to have control over something is to have causal power over it, and to have causal power over something is for its existence or nature be causally explained by factors relevantly associated with the agent. (What counts as "relevantly associated here" here can get complicated, particularly with regard to Billy's control over his motives.[15] But bracket this for now.) Extending the analysis, we can say that the events of his intentionally

15. To anticipate the next section, it may be thought that in asking whether Billy had the normal degree of control over his motives, we are asking really asking whether Billy's breaking of the window was attributable to him in a normative sense, as embodying his normative self-conception. But this is not necessarily so. Such is evident, for instance, in the law, wherein the question whether Billy's process of indoctrination had exculpating features (arguably) can and should be kept distinct from the question whether his motives were authoritatively representative of him. (If all this seems a bit much for rock-throwing, suppose the offense were more serious.)

breaking the window, or having certain motivations, are likewise attributable to him to the extent that they are likewise subject to his causal power, again in virtue of admitting of the right kind of causal explanations.

For something to be attributable to you in this sense is for you to have the right kind of causal power over it. As Billy's case suggests, such attributability is plausibly as a kind of quasi-juridical accountability. It is somewhat common to identify moral responsibility with such accountability, and very common to identify free agency with the capacity to be morally responsible. This makes it unsurprising that many influential writers on free will accept a causal view of the attributability that distinguishes free actions from merely random ones. Hume clearly does, for instance, when he argues that "where [actions] proceed not from some cause in the characters and dispositions of the person who performed them, they infix not themselves upon him, and can neither redound to his honor, if good, nor infamy, if evil."[16] So does Roderick Chisholm, when he asserts that "if the cause [of an action] was in some state or event for which the man himself was not responsible, then he was not responsible for what we have mistakenly been calling *his* act."[17] Hume and Chisholm disagree only with respect to the potential causal *explanantia* of free actions: Hume limits them to the agent's internal states and dispositions—paradigmatically, beliefs and desires—while Chisholm argues that actions can be caused by agents conceived as enduring substances, rather than as complexes of states or events.

Now, there is a familiar line of thought leading to incompatibilism from a causal conception of attributability. If Billy's throw breaks the window because Suzy knocked his rock into it, we deny that the breaking of the window is attributable to him because it results from factors outside of his causal power, and so beyond his control. Similarly, if he chose to break the

16. *A Treatise of Human Nature*, §2.3.2.

17. "Human Freedom and the Self," p. 27, my emphasis.

window because of Black's remote neural manipulation, we likewise deny that his *choice* was attributable to him because it was, similarly, determined by factors beyond his control. And if he was motivated to break the window because of Black's anarchist indoctrination, we might once more deny that his *motives* were attributable to him because they were, yet again, determined by factors beyond his control. Convinced that what makes the difference in all three cases is determination by factors beyond one's control, we might see no principled basis for changing our answer if the factors outside the agent's control are simply those of history unfolding in its normal way, in accordance with deterministic laws of nature.[18] So in order for Billy's breaking of the window to *really* be up to him, it must be undetermined not only by any of the familiar sorts of exculpating factors above, but also by anything else.

But if the line of thought leading from a causal conception of attributability to incompatibilism is familiar, so is the one leading from incompatibilism to a causal conception of self-creation. It was once common to argue that an action could be attributable to you only if it was causally *determined* by the relevant factors. This argument is not itself the problem. Even if antecedent conditions do not causally determine which of a range of choices you will make, it is plausible that in some cases, any of the choices open to you could have the right kind of causal etiology to count as under your control. The problem is that it may yet not be under your control *which* of these possible choices you actually end up making.

While versions of this challenge have actually been developed in detail in the recent literature, we can get at the basic idea by adapting Nagel's core argument against the possibility of autonomy.[19] While Nagel's argument is explicitly targeted against normative conceptions

18. Derk Pereboom defends the canonical contemporary version of this argument (which proceeds very similarly to the one sketched above) in "Determinism *al Dente*," along with many variations and elaborations of it in later work.

19. But see, e.g. the critical responses to Robert Kane's defense of libertarianism in "Responsibility, Luck, and

attributability, extending it to the present case will help bring out the important structural similarities of the cases. Nagel writes:

> The intuitive idea of autonomy includes conflicting elements, which imply that it both is and is not a way of explaining why an action was done. […] When someone makes an autonomous choice such as whether to accept a job, and there are reasons on both sides of the issue, we are supposed to be able to explain what he did by pointing to his reasons for accepting it. But we could equally have explained his refusing the job, if he had refused, by referring to the reasons on the other side—and he could have refused for those other reasons: that is the essential claim of autonomy. It applies even if one choice is significantly more reasonable than the other. Bad reasons are reasons too.
>
> Intentional explanation, if there is such a thing, can explain either choice in terms of the appropriate reasons, since either choice would be intelligible if it occurred. But for this very reason it cannot explain why the person accepted the job for the reasons in favor rather than refusing it for the reasons against.[20]

Again, bracket—for now—Nagel's focus on the idea of "intentional explanation" and imagine the case concretely. Suppose that on the one hand, you are tired of living in your parents' basement, and accepting the job would enable you to get a place of your own. On the other, you are spiteful toward your parents, and refusing the job would enable you to vex them. Now suppose that your process of weighing these reasons in deliberation is causally undetermined in such a way that, prior to coming to a conclusion in favor of one or the other of these options, it is perfectly possible that you will make either choice. That is, whichever choice you make, you'll experience it as a perfectly normal outcome of deliberation; your accepting or rejecting the job will be intelligible to you or anyone else as a perfectly normal intentional action, manifesting a perfectly normal degree of control. What's so incoherent about that?

Well, nothing, yet. The problem is that what you *don't* as yet have control over is whether the nondeterministic causal factors that undergird your deliberation—working "under the hood," so to speak—operate in such a way that you come to one conclusion rather than the other. In fact, it's hard to see how what your having control over them would even involve, since it's

---

Chance," which argue along similar lines that Kane's form of libertarianism cannot secure a form of control not also available to compatibilists.

20. *The View from Nowhere*, pp. 115-116.

presumably possible to deliberate without even knowing they're there. This point can easily be put in terms of causal explanation. Plausibly, the choice you make (whichever one it turns out to be) is indeed attributable to you because it has the right kind of causal explanation—in terms of the way your motives shape your deliberation, say. But what is *not* attributable to you is the fact that your motives shaped your deliberation such that it had the outcome it did. For that there may be no explanation at all.

Whether this is in fact a problem depends on what the point of origination is supposed to be. But according to the line of thought I've sketched above, the point of origination is supposed to be responsibility: if something was determined by factors outside your control, you are not responsible for it. Now, admittedly, it's true with regard to the job case that the fact that your deliberation went the way it did was not determined by factors outside your control. But it certainly wasn't determined by factors *within* your control, either, for the obvious reason that it wasn't determined by any factors at all. Notoriously, this forces the question: if you are not responsible for anything if it is determined by factors beyond your control, what condition on responsibility has gone unmet? Is it the negative condition of the absence of external determination, or a positive condition of control? It is hard to see how the former of these could make a difference.[21]

But if a mere absence of external determination indeed isn't enough for autonomy, we're in trouble. In this case, autonomy would require you to have causal power not only over your choices, but also over the inner causes of those choices. Again, I have been assuming in this section that to have causal power over something is for it to have the right kind of causal

21. Much of the discussion of this point in the free will literature has been in relation to thought experiments (starting with one proposed by Harry Frankfurt) purporting to specify conditions under which an agent clearly lacked the ability to do otherwise (for reasons that having nothing to do with determinism) but was nevertheless clearly morally responsible for some misdeed. (For a recent critical summary in support of the same conclusion I draw, see John Martin Fischer, "The Frankfurt Cases: The Moral of the Stories.")

explanation, in terms of certain factors relevantly associated with the agent. So the operation of the inner causes of those choices must themselves have a causal explanation of the right kind, in terms of still *other* relevant agential factors. (In order to avoid an infinite multiplication of motivational states, it will be easier to imagine the "relevant agential factors" as more like acts of will, along traditional incompatibilist lines.) Now, unless we want to run into the same problem at a higher level, these other factors—whatever they are—can't just operate non-deterministically: rather, *their* operation must be under your causal power *too*. Which means they too must have a causal explanation of the right kind, which means there must be yet more relevant agential factors, which means more things that must be under your causal power and more causal explanations of the right kind, and so on to infinity. To avoid the regress we can have two options. We might try to formulate a notion of control that still consists in the exercise of a causal power but does not imply the availability of a causal explanation. (I take something along these lines to be Chisholm's motivation for positing agent-causation.) But, as Nagel put it, this seems to be giving a name to a mystery. The only other option is to claim that some of the relevant agential factors are causally explained by one another—to posit a finite loop, instead of an infinite line. But, since nothing can causally explain another without causing it, this would amount to claiming that some causes causally depend on their effects. Presumably this is metaphysically impossible. Even if it isn't, it's at least as mysterious as the other option.

Let's review. Autonomy in the causal sense consists in the capacity to make choices that ate attributable to you in that they are under your control, but originate with you in that they are not determined by any causes outside of your control. But if autonomy in the causal sense is important, it is evidently not so because it is important to choose in the mere absence of determining causes, but because it is important to have ultimate control over whatever causes your choices *do* have. But in order to actually exercise such control as a finite agent, you would

have to be capable of creating yourself, in the sense of being an agent-cause or a *causa sui*. Whether self-creation in either form is intelligible is dubious at best.

It's worth reiterating that nothing I've said in this section is meant to be at all original: all I've done is rehearse arguments that anyone who has read anything about free will is likely to have seen before. I've rehearsed them here because I wanted to make explicit that both much of the appeal of the causal conception of origination, and much dialectical pressure leading from that conception to an evidently impossible requirement of self-creation, come from the same place: the causal, quasi-juridical conception of attributability. If this were the only conception of attributability in our repertoire, and if the concern for autonomy was just a concern for ultimate control, then it seems to me that the idea of autonomy really would be hopeless. But this isn't the only conception of attributability in our repertoire. As I now want to argue, we'll be better able to make sense of what might be so compelling about autonomy in the first place by interpreting it in terms of a different one.

4 The normative conception

Attributability in the causal, quasi-juridical sense is grounded in attitudes that could equally intelligibly be taken toward someone else's action as toward one's own. But, as many recent philosophers have stressed, sometimes we care about attributability in a different, essentially first-personal sense, grounded in attitudes agents take toward their own actions in practical reasoning. To get a basic sense of the difference between these approaches, take Martin Luther at the Diet of Worms, refusing to recant—"here I stand, I can do no other." This case is often deployed in the free will literature as evidence in favor of compatibilism, by showing that experiencing an action as necessary is compatible with experiencing it as something you do freely. But Luther's experience of the necessity of his action is not, primarily, a matter of *causal* necessity. "When

Luther says he *can* do nothing else," Nagel observes, "he is referring to the normative irresistibility of his reasons, not their causal power."[22] He is not predicting how he is likely to behave, given the condition of his internal states and dispositions.

In fact, causally speaking it may indeed be possible for Luther to do something else. Suppose (as some libertarians have proposed) that in some stressful situations, the effects of quantum indeterminacy in the brain are magnified so as to make macro-level processes of decision-making nondeterministic.[23] And suppose this is just what happens to Luther. Understandably troubled by the threat of being branded an outlaw and heretic, Luther may be terrified, tired, and wish on a kind of raw animal level that he could just go home. But suppose he knows that these emotions do not register anything that actually matters about the case, or, really, what matters to him; he is in absolutely no doubt about either. So he gives no weight to his feelings in practical reasoning. But they are there all the same, and there so strongly that, up to the moment of his decision, the objective chance of his breaking down and recanting precisely matches that of his sticking to his principles.

But this would not falsify his declaration. Again, "I can do no other" is a decision, not a prediction. It expresses a conclusion reached in practical reasoning, on the basis of the evaluative judgment that the reasons in favor of refusing trump the reasons in recanting. Luther would experience refusing as something he is *actively* doing, and the recanting as something he is *failing to resist* doing. That is, it is conceivable that he may break down and fail to act on this conclusion, but if he does he will experience his action as independent of his practical evaluations and, hence, against his will. When Luther says *he* can do no other, that is, he is referring to himself under a

22. *The View from Nowhere*, p. 117n3. To be clear, I aim not claiming the example generalizes to all cases of practical necessity. In other cases, the recognition that one is indeed psychologically incapable (in a largely but not exclusively causal sense) of a certain type of action can play the important role of mooting deliberative questions that may otherwise have been very fraught.

23. Cf. e.g. Robert Kane, *The Significance of Free Will*, pp. 128-130.

very specific description—as an agent essentially identified with certain normative commitments. He may, as an empirical matter, fail to adhere to these commitments, but if he did he would *ipso facto* fail to act as *himself*, in the deepest sense.

So Luther is referring to himself in a *normative* sense, not a causal one. More generally, a choice is attributable to you in a normative sense just in case it is chosen according to the norms to which you hold yourself fundamentally answerable in practical reasoning. Naturally, defenders of different versions of the normative conception will understand what it means for a choice to be "made according to the norms with which you hold fundamentally answerable in practical reasoning" in different ways.[24] But what is common to any version of the normative conception is that it is at most a necessary condition of a choice's being attributable to you that it admit of the right kind of causal explanation. That is, while Luther's refusal must have been made it in light of an evaluation that it was, indeed, appropriate with respect to these norms, it could only count as attributable to him in a normative sense if the evaluation was correct.[25] So attributability is still a matter of explanation, but in the normative case there must be a *justifying* explanation, above and beyond a causal one. This explanation does not purport to explain how Luther's action was *produced*, given his constitution and environment. It rather purports to show how it was *intelligible*,

24. For important defenses of the normative conception of character, see e.g. Bernard Williams, "Persons, Character, and Morality," Harry Frankfurt, "Identification and Wholeheartedness," Gary Watson, "Free Agency," Christine Korsgaard, *The Sources of Normativity*, ch. 3; Michael Bratman, *Structures of Agency* (esp. chs. 2, 4, 7, and 8); and Charles Taylor, "What is Human Agency?" In §63 of *A Theory of Justice*, Rawls presents an abstract but prototypical version of a normative conception of character, in which the norms in question constitute the agent's "plan of life," and traces the normative conception in general to early twentieth-century idealists and pragmatists, like Bradley, Royce, Dewey, and Perry. Rawls writes: "Here I adapt Royce's thought that a person may be regarded as a human life lived according to a plan. For Royce an individual says who he is by describing his purposes and causes, what he intends to do in his life." Rawls argues that it is only in virtue of the content of these plans—and then only insofar as the plans are rational—that persons can be intelligibly such that things can be good for them. In a footnote, he makes the constitutive role of these norms even more explicit, adding that "Royce uses the notion of a plan to characterize the coherent, systematic purposes of the individual, what makes him a conscious, unified moral person." (*A Theory of Justice*, p. 358)

25. Here it is important to keep in mind that we can be very permissive about what counts as an evaluation that some action is appropriate with respect to a norm. Huck Finn's failure to turn Jim in may be attributable to him in the normative sense even if his inability to tell the truth when asked reflects the implicit awareness that to do so would be to grossly betray a friend—even if he only interprets his attitudes as such in retrospect.

as part of a life embodying a certain set of values. As such, its *explanatia* are facts about the *content* of Luther's values, not facts about his internal states or dispositional one.[26] Asked why Luther refused to recant given the physical, social, and psychological costs, one might give a causal, psychological explanation, citing his relevant motives and features of his constitution in virtue of which those motives effectively moved him to action. But someone might reasonably reply: "Yes, yes, I know all *that*, but who cares? What I want to know is why it *made sense* for him to recant." Here one might begin with the obvious "well, you *know* he could do no other, right?"—and then explain why he could do no other by citing the considerations that justified his choice from his perspective ("I am bound by the Scriptures I have quoted and my conscience is captive to the Word of God," and so on), perhaps relating them to his larger system of values so as to show how, relative to that system, those considerations weighed so heavily. . This explanation does not purport to explain how Luther's action was *produced*, given his constitution and environment. It rather purports to show how it was *intelligible*, as part of a life embodying a certain set of values. As such, its *explanatia* are facts about the *content* of Luther's values. It appeals on facts about his internal states or dispositions at most indirectly, insofar as these facts determine the content his values have.

The distinction between causal and normative conceptions of attributability is familiar. But what may be less familiar is that we can draw the same distinction with respect to origination. In fact, Luther's "I can do no other" might plausibly be taken to imply that he *didn't* think of himself as the originator of his choice in a normative sense. Even if the antecedent didn't determine Luther's empirical behavior, they still determined what Luther *qua* normative self-conception would choose, so to speak. Recall that when I first described origination, in §2, I

26. As I noted in fn. 14, my use the term "justifying" here should not be taken to mean that an attributable choice must necessarily justifiable to anyone other than yourself.

offered that a "a choice originates with you if all of the facts about your circumstances and yourself—antecedent to and independent of the choice itself—leave it open to you whether or not to make it." This should have read like a pretty direct, basic statement of the kind of origination incompatibilists defend. But we can now see that "you" admits of multiple interpretations here: it could refer to "you" *qua* locus of empirical causes, but it could also refer to "you" *qua* normative self-conception. What we've seen in the Luther case is that the former kind of origination doesn't entail the latter, and it should be pretty clear from the last three chapters that there's no entailment in the other direction, either. To go back to the first chapter: even if it's causally determined that Sartre's pupil will leave his mother to fight, there may be no fact of the matter—relative to everything in principle accessible to him in deliberation—whether this is in fact the choice the norms constitutive of his self-conception call for.

As we saw in the last section, being the causal originator of your choices can seem important to many people because it can seem important to have ultimate control over their causes. Without such control, one might understandably think, what to do would not really be up to *you*. It would rather be up to external causes, or to chance. Extending the analogy suggests that if it important to be the *normative* originator of your choices, it is because it is important to have ultimate *authority* over their standards. Without such authority, one might no less understandably think, what to do would not really be up to you, either. It would rather be up to external normative requirements, or to nothing at all.

Some people find the idea of ultimate control obviously important and do not need to be argued into it. Others find it confused, if not embarrassing or worse, and are equally sure of their convictions. I suspect something comparable is true of ultimate authority. Luther certainly had no interest in it. It is doubtful that the possibility of ultimate first-personal normative authority would even be intelligible in his moral framework; it certainly wouldn't be worth wanting. For

him all authority rested with God. Needless to say, more or less secular analogues are common in philosophy: Plato and Spinoza are the prime examples, along with the many contemporary defenders of orthonomy. But if I have been right in the two previous chapters, about love and blame, then denying the intelligibility or importance of ultimate first-personal normative authority really would render unintelligible much of the interest we take in ourselves and others as particular individuals. (Whether it would similarly render unintelligible the attitudes of notoriously intense ancient, late medieval, or early modern mystics is another question. They may simply have seen the world and our place in it differently.)

Of course, the discussion of love—and less directly the one of blame—did not so much appeal to ultimate first-personal normative authority as such, but rather to a specific account of the relation between the content of norms to which an agent is committed and the choices the agent takes to be appropriate with respect to them. This suggests a respect in which the analogy between the causal and normative cases breaks down. In the causal case, it is possible to get a more or less intuitive grasp on why ultimate control is supposed to matter prior to formulating a reasonably systematic account of how such control might actually be realized in human agency. Rather, all we need to get the more or less intuitive grasp is a version of the quasi-juridical conception of attributability and a sense of how the conditions under which such attributability is undermined might constitute a slippery slope. I think the normative case runs in the opposite direction. (This is yet another reason why the appeal of autonomy can be so elusive.) The best way to appreciate the importance of ultimate first-personal normative authority is to think through what we would have to be like to have it, and explore the implications that our being this way would have for our attitudes towards ourselves and others. Origination in the normative sense is compelling *because* of the kind of agency necessary to adequately realize it, not despite it.

So before I survey, as a kind of summing-up of the dissertation as a whole, why the normative sense of origination might have been the one that mattered all along, it's first necessary to finish drawing the analogy and trace the normative version of the dialectic from origination to self-creation. This is just a recapitulation of the argument of the first chapter, which aimed to reconstruct and defend a version of Sartre's claim that his pupil's choice could be rationally undetermined without being arbitrary. Start by imagining Sartre's pupil in terms that precisely parallel the person Nagel described as faced with a choice whether to accept or reject a job. Here, too, there are reasons on both sides, and we can assume that Sartre's pupil would be more or less intelligible in choosing either of the alternatives on the table.

We saw in the causal case that stopping here would give us a very simple way of accounting for origination. This was the "leeway incompatibilist" route: to claim that the origination relevant to autonomy simply consists in the presence of indeterminism in normal deliberative processes, of the sort otherwise identical to those defended by compatibilists. The leeway incompatibilist answer has the benefit of being unproblematically coherent, but the drawback of being inadequate to the intuitive importance of autonomy in a way that Nagel's objection captured. Nagel's objection was that either of a pair of possible choices might both be explicable as done for reasons, and so attributable in a basic sense, but if they were, there could not also be an explanation why one chose to act on the reasons one did, given the availability of the other set. By interpreting these explanations as the kind of causal explanations characteristic of something's being under an agent's control, we saw how naturally the argument applies to the causal case: there may be two causal processes left upon by antecedent conditions, each constitutive of your exercising your control in a certain way, without it being under your control *which* of these processes actually takes place. But it is hard to see how origination that lacked such higher-order control is supposed to matter.

On the normative level, precisely parallel accounts are common. We might say that norms to which the pupil is committed leave both staying with his mother and leaving to fight "rationally eligible," but do not specify weighing principles or other standards for comparative evaluation; in deciding which set of reasons to act on, all the pupil can do is plump.[27] So antecedent conditions fail to determine what the pupil will do *qua* normative self-conception, just as in the previous case they failed to determine what he would do *qua* locus of empirical causes. So here, too, the pupil's choice might technically count as originating with him, but only in a very shallow sense. The problem continues to be the one Nagel identified, just transposed to a different kind of explanation. Each choice may admit of a basic kind of justifying explanation, in terms of the reasons that render it rationally eligible. But this basic kind of justifying explanation necessarily cannot be a justifying explanation of why the pupil chose to act on the reasons he did. And just as people worried that determinism might prevent them from being more than superficially accountable for their choices might understandably balk at not having control over which undetermined possibility occurred, so might a Sartrean, concerned to live in authentic recognition of one's own role in defining oneself, likewise object to the denial that he could not intelligibly take responsibility for his choice as putatively expressive of what matters most to him.

Now, if it is still to be open to the pupil to choose *either* option, it must be possible *both* that there could be an all-things-considered explanation why he ought to have stayed with his mother, in the event of his making that choice, *and* an all-things-considered explanation why he ought to have left to fight, in the event of his making that one. Since it is presumably impossible that a single possible world should include all-things-considered explanations, of precisely the same kind, of both the truth of a proposition and the truth of its negation, the only way for a choice to both meaningfully originate with you and be meaningfully attributable to you is for its

27. Cf. Joseph Raz, "Incommensurability and Agency."

explanation to partly depend on the very fact that it is made. This applies no less for causal explanations than normative ones; it reflects the basic interdependence between the ideas of origination, attributability, and self-creation, conceived at a high level of generality. But while in the causal case the requisite self-creation is, again, metaphysically impossible—causal *explanantia* cause their *explananda*, and causes cannot depend on their effects—in the normative case, as I argued in the first chapter, it isn't.

I won't repeat these arguments here, but it's worth recalling the concrete picture of Sartre's pupil they imply. To begin with, the pupil may safely be imagined as anguished. It's not only that there are reasons on both sides: it's that there are pressing *reasons* on both sides, to the point that each alternative is such that choosing against it would be a proper occasion for remorse and a deep sense of loss. This feature of the case has led some philosophers to interpret it as a dilemma. But I do not think Sartre meant the pupil's anguish to be understood as the same kind of thing as the truly paralyzing sense of inner conflict characteristic of truly tragic dilemmas, like Sophie's or arguably Agamemnon's. In a truly tragic dilemma, it's obvious what the weights of the relevant considerations are: they're categorical on both sides, and the problem is that it's therefore impossible choose correctly. The pupil's case is the opposite: the problem is that it *is* possible to choose correctly (even though doing so will be painful), but it's anything *but* obvious what the weights of the relevant considerations are. What makes his choice an occasion for anguish is that he will regret it terribly if he chooses incorrectly, and that he is forced to choose in the absence of anything like a principled, rational basis for identifying the correct choice in advance. He needs to choose, but the most he could possibly have to go on is a primitive, inchoate feeling that one course of action somehow embodies how he aims to live in a way that the other doesn't. This feeling is explanatorily impotent on its own: he will only be in a position to explain or justify its authority after he has lived with his choice long enough, and coherently

(or "authentically") enough, to work out a more developed sense of self in its wake. Had he felt and chosen otherwise, and lived with *that* choice, it, too, might have turned out to correct (with "*might*" being the operative word here).

Now, suppose the pupil decides to leave his mother to fight, and further suppose, optimistically, that that through doing so he comes to both live as an adventurer and patriot and understand himself as one. From the evaluative point of view that he thereby comes to inhabit, it will be perfectly possible for him to provide a justifying explanation of his decision to leave his mother rather than stay with her. Admittedly, that explanation was not accessible to him *in advance*, and couldn't have been in principle. But who said it had to be? To intelligibly regard his choice as free, the pupil will have to regard it as attributable to him in the sense of being (contrastively) explained by who he is. So the explanation of the pupil's choice must be accessible to him in advance only if he must regard who he is—his self or character—as *itself* something whose nature must be somehow be fully accessible or extant in advance of his actions. But, again, why must he regard himself as such? On a causal view, of course, we would have a perfectly good argument here: he would have to regard his attributable actions as caused by himself or his internal characteristics, and causes must precede their effects. But we have just seen that it is precisely from the pupil's first-person perspective in practical reasoning that he need not conceive himself in such a way, and so from *this* perspective the requirement that the pupil's character be accessible or extant in advance is unmotivated. On the contrary, it's perfectly natural to say that from a first-person perspective, the pupil *necessarily* doesn't view who he is as accessible to him in advance, precisely because he views who he is as something he's in an ongoing process of determining. This shows how the origination and attributability conditions can be satisfied together: the pupil lacks a prior justifying explanation of his action not because there is no such explanation, but rather because the features of himself that support it themselves depend on a

sequence of judgments and actions that include (but go beyond) the very action he is about to choose. Relative to the *antecedent* state of his character, the pupil could have freely and attributably done otherwise.

Let's return to Nagel's assessment of the "logical goal of [the] ambitions [implicit in our ordinary idea of autonomy]"— "to be really free we would have to act from a standpoint completely outside ourselves, choosing everything about ourselves, including all our principles of choice—creating ourselves from nothing, so to speak." We can now see that on the normative conception of self-creation, this is partly true and partly false. The freedom of Sartre's pupil does, in a sense, require him to choose at least many of the most fundamental things about himself, including his principles of choice—at any rate, it requires him to reason and act according to norms whose content it is fundamentally up to him to determine through antecedently undetermined choices. But it does not require him to act from a standpoint completely outside himself—although I suppose it would if he had to create himself *all at once*, which really would be impossible. It just requires him to act from a standpoint that is never fully complete or determinate, because it's the standpoint of someone always in the process of becoming who he is.

5 The importance of self-creation

One of the things the case of Sartre's pupil brings home is how far the ideals of ultimate control and ultimate authority can come apart. While it is indeed up to him and only him to define who he is through his choices, he doesn't actually have all that much *control* over who he is, because he doesn't have all that much control which choices actually he ends up making. And note that he lacks this control for reasons that have very little to do with the metaphysics of agency. Give him the most metaphysically extravagant form of libertarian freedom you can think of: so long as it falls short of turning him into a god, his *external occasions* for choice will still be largely out of his

control, and these play just as much of a role in setting the boundaries of who he can make himself into as the internal causes of his choices.

We should therefore expect the appeal of normative autonomy to come from a very different direction than that of its causal counterpart. In fact, normative autonomy offers no comfort to people whose unease over determinism stem from a concern for quasi-juridical accountability. But it is nevertheless essential to other phenomena equally prominently associated with free will. It follows from the two previous chapters that attributability and origination, in their normative forms, are required for two of the richest and most important Strawsonian reactive attitudes: angry blame and interpersonal love. This in itself both vindicates and defuses the traditionally incompatibilist conviction that a central form of human freedom indeed requires a much stronger form of self-creation than either hierarchical or orthonomy compatibilist views would imply. In insisting on the importance of self-creation, Kane and incompatibilists like him were right all along.

Taken together, the arguments of these chapters have a broader lesson to teach, about the nature of individuality and freedom of the individual. In order to respond to someone as a particular individual, it is necessary to take an interest in that person as having a specific personality and character, constituted by a set of fundamental values or commitments—and not ones that would necessarily be affirmed or even tolerated from an objective or impersonal standpoint. But it is also necessary to implicitly regard those commitments as ones that necessarily cannot be individuated in abstraction from the individual's ongoing sequence of particular judgments and actions. If people did not act from characters they were in the process of freely creating—if it were in principle possible to pin people down, to definitively specify their essential commitments in abstraction from their particular histories, as fixed things—it would not

be possible to fully regard them as individuals, in a way that would undermine the intelligibility of many of our central attitudes and practices.

Explaining the discomfort many people seem to feel with the possibility of being predictable according to deterministic laws of nature, Kane suggests that "most people want to ascribe a uniqueness to themselves that would make it impossible for others to treat them as types, subsuming all their behavior under general laws."

> They want to say, "Don't type me. Pay attention to *me* and not to your physiological, psychological, or social formulae, because I will surprise you, no matter how comprehensive your knowledge is. To deal with me as a person, you must wait to see what I will do and react accordingly."[28]

One might reasonably wonder how causal origination or self-creation is supposed to help. Are you treating someone less as an instance of a type, and more as an individual, if the physiological, psychological, and social formulae are probabilistic instead of deterministic? (Does recognition of individuality increase as ideal credences in outcomes approach .5?) But if the worry Kane describes is interpreted as a concern for normative self-creation, it is both understandable and apt.

We can see its importance most vividly with respect to love. If it is important to be valued as a unique individual anywhere, it is important to be valued as such by a lover. In the second chapter, we saw how Frankfurt's attempt to explain how beloveds could be valued as unique individuals was inadequate because he denied that there were reasons for love, and more specifically because he denied that love was a appreciative response to the beloved's essential character. Catherine could not love Heathcliff for what his soul was made of. In this chapter, we saw how Frankfurt's hierarchical view of freedom can provoke intuitive dissatisfaction in part because it conceives an agent's essential character as constituted by a given complex of psychological states, something one does not create for oneself but comes to have in virtue of

28. *The Significance of Free Will*, p. 86.

"nature and the world." Viewing these aspects of Frankfurt's thought side by side, we can see how his account of love could *only* be what it is. For Frankfurt, to be loved for your character would be to be loved for something that would make you too replaceable: an arrangement of psychological states that could be equivalently instantiated, and equivalently lovable, in someone else. Velleman thought he could avoid this problem by construing the beloved's essential character as something that transcended their empirical psychological properties, arguing that interpersonal love was a response to the value of the beloved's bare Kantian personhood. But that response, I argued, goes too far in the opposite direction. Catherine's love for Heathcliff is not an indiscriminate openness to what makes him a *person*; it is a discriminating and insistent attraction to what makes him specifically *Heathcliff*.

In order to avoid these extremes, we need a conception of character as something that transcends a person's existing characteristics while remaining specific to that person. We need the same kind of intermediate conception, I argued, for an adequate account of blame. The uneasy combination of emotional vulnerability and defensiveness distinctive to righteous anger requires that there be a real question whether an offense was characteristic of the agent who performed it, but also that the question be in principle unanswerable prior to the expression of anger and its acceptance or rejection by the agent. The improvisational model, distinctively, supports just such a conception.

Once we recognize the need for this intermediate conception, we can see how appeals to orthonomy are responsive to a source of deep intuitive dissatisfaction with hierarchical compatibilist theories but misaddress it. C. A. Campbell writes, for instance:

> What [the experience of free agency] implies—and it seems to me to be an implication of cardinal importance for any theory of the self that aims at being more than superficial—is that the nature of the self is for itself something more than just its character as so far formed. The 'nature' of the self and what we commonly call the 'character' of the self are by no means the same thing, and it is utterly vital that they should not be confused. The 'nature' of the self comprehends, but is not without remainder reducible to, its 'character'; it must, if we are to be true to the testimony of our

> experience of it, be taken as including *also* the authentic creative power of fashioning and re-fashioning 'character'.[29]

Part of what it is to regard yourself as free from the perspective of practical reasoning, Campbell claims, is to regard yourself as free *of* your existing motivations. What "reflective human beings want," Nagel writes in a similar spirit, is "to be able to stand back from the motives and reasons and values that influence their choices, and submit to them only if they are acceptable." (127) As I noted in §1, Frankfurt's hierarchical view is motivated by the need to account for such reflective distance. As I also noted in that section, however, it is controversial whether Frankfurt succeeds in this. It no less possible to reflectively distance yourself from your higher-order motives as from your first-order ones. As many of Frankfurt's critics have argued, avoiding an infinite justificatory regress requires, ultimately, endorsing your will as appropriate with respect to standards that do not depend on it. And note that such a conclusion, and the motivation behind it, is not unique to compatibilists. Campbell himself turns out to argue that the only occasions for free choice are choices between what you most strongly desire (where this is strictly a function of your "character as so far formed") and what you morally ought to do (which Campbell tends to call, rather quaintly, your "duty"). Thus, the "authentic creative power of fashioning and re-fashioning character" turns out to be the power to fit your character into the moral mold (which makes one wonder whether "creative" was really the right word).

And so we are led to orthonomy. When Nagel endorses orthonomy as a partial reconciliation of our intuitive idea of freedom with an objective view of ourselves, he does so in self-conscious debt to Kant. There is, Kant famously argued, only one way to reconcile our freedom from our existing motivations with the possibility of our acting for reasons, and that is to conceive of ourselves as free only when acting for reasons that *any* agent acting "under the idea of

29. *On Selfhood and Godhood*, p. 177.

freedom" would act for, regardless of whatever contingent motivations such an agent might happen to have. But we can now see how Nagel's Kantian account of freedom goes too far in precisely the same way that Velleman's Kantian account of love does. The phenomenon of reflective distance indeed requires fundamentally identifying yourself in practical reasoning with norms whose content necessarily depends on more than your present motivations alone. But one might think, as Bernard Williams has argued throughout his work, that this *cannot* mean fundamentally identifying yourself with the essentially impersonal norms of morality, on pain of ceasing to regard yourself as a particular individual, with a distinct personality and character of your own. The question of whether Williams is right here is obviously beyond the scope of this chapter. But if he is, it follows that the conception of individuality in the Kantian view of freedom is just as superficial as the one in the Kantian view of love, and for precisely the same reasons. In that case, both the hierarchical and orthonomy views fail as accounts of freedom of the individual. The nature of the self needs to be for itself more than just its character as so far formed, but it can't be *characterless*. We need a middle ground, and to get it, we need origination, and ultimately self-creation, just like the incompatibilists thought.

6 Conclusion

I began the chapter by noting the conviction many people have that the capacity to be the originator of your acts is vital to the nature and importance of free agency, and the failure on the part of compatibilists to adequately address this conviction. I have argued that compatibilists should offer a compromise. They should accept that origination is indeed vital to the nature and importance of free agency, in a way they have thus far denied. But they should ask, in return, that their counterparts reconsider whether the capacity at issue is really as metaphysically tendentious as they thought.

To this end, I have traced how the basic, intuitive idea of autonomy as origination admits of two natural interpretations, corresponding to two different conceptions of the self as a source of agency, and each yielding structurally similar but substantively different conceptions of autonomy as self-creation. Both of these conceptions are responsive to old and powerful sources of attraction to the idea of self-creation, but they differ strikingly with respect to the natures of their respective attractions. The causal conception reflects an ideal of autonomy as *ultimate control*, grounded in the value of a robust form of accountability. The normative conception, on the other hand, reflects an ideal of autonomy as *ultimate first-personal normative authority*, grounded in the value of a robust form of individuality.

Whether or not I have succeeded in actually motivating a compromise, therefore, I at least hope to have forced attention to the question: what is it about self-creation that makes it vital to the nature and importance of free agency? Is it a matter of control and accountability, or first-personal authority and individuality? Those who prefer the second answer will have reason to accept the compromise on offer; those who prefer the first won't. Naturally there is no reason to expect consensus: I suspect this is one of those issues that mark basic divisions among philosophical temperaments. But I think our understanding of the free will problem will be advanced if we recognize it as turning, fundamentally, on this broadly ethical question, to as much or greater an extent as on questions about the metaphysics of agency.

I want to finish, in this spirit, by gesturing at some reasons for thinking that the divide between the ideals of ultimate control and ultimate authority is very deep indeed. In the question-and-answer session following *Existentialism is a Humanism*, the writer Pierre Naville pressed Sartre with a long and challenging series of objections. These include the following, which could be repeated verbatim against the view I have proposed:

> It is not true that man has freedom of choice in the sense that such choice allows him to endow his actions with a meaning that they would not have had without it. It is not enough to say that men can fight for freedom without knowing they are doing so; or then, if we were to attribute such recognition in its full meaning, that would mean that men can commit themselves to, and fight for, a cause that dominates them, which is to say an act within a context that is beyond them, and not only in their own terms. For in the end, if a man fights for freedom without knowing or expressly formulating for himself in what way, and for what purpose, he is fighting, that means his actions will bring about a series of consequences that would insinuate themselves into a causal web, all the facets of which he would not totally grasp, but which would nonetheless delimit his actions and give them a meaning in terms of other people's actions—not only those of other men, but of the natural environment in which such men act.[30]

For the most part I have presented my view as broadly descriptive—as an analytical reconstruction of ethical attitudes and practices when conceived in terms that would answer to their evident importance. But the implication that Naville ascribes to Sartre, and which my view shares, is undeniably revisionary—if not to the ethical attitudes and practices themselves (which may be silent on the matter), at least to the low-level, informal theorizing that occurs at the level of abstraction immediately above them. It is certainly completely at odds with what many philosophers—and above all most incompatibilists—view as the essence of autonomous agency. Even compatibilists frequently insist that while you may not have complete control over whether your projects work out the way you want them to, insofar as you are capable of acting and reasoning autonomously, you are at least capable of guaranteeing that your choices about how to pursue those projects are always appropriate by the standards to which you can reasonably hold yourself. My view that even this internal control is not guaranteed. No matter how conscientiously Sartre's pupil might go about making his choice, whether he turns out to be justified by his own standards in making it depends on the choices that the world, and other people, enable him to make down the line.

Nevertheless, I do not regard this consequence as much of a problem. Essential dependence on contingency is just the price of ultimate first-personal authority. As we saw in the first chapter, and as Frankfurt reinforced in the quotation at the beginning of this one, the idea of

30. *Existentialism is a Humanism*, p. 64.

ultimate first-personal authority would be unintelligible if it amounted to defining who you were by sheer, unconstrained stipulation: "A person's will is real only if its character is not absolutely up to him." The only way I can even begin to see of reconciling the possibility of such constraint with the possibility of one's commitments ultimately depending on one's choices is the one I've defended here. But if self-creation is only intelligible as a process that essentially extends before and after a choice, contingency inevitably creeps in.

But this is just to give a reading of a famous passage that already constitutes the best and deepest reply to Naville, though it wasn't written by Sartre. It wasn't even addressed to Naville, although it reads like it could have been:

> To insist on such a conception of rationality, moreover, would, apart from other kinds of absurdity, suggest a large falsehood: that we might, if we conducted ourselves clear-headedly enough, entirely detach ourselves from the unintentional aspects of our actions, relegating their costs to, so to speak, the insurance fund, and yet still retain our identity and character as agents. One's history as an agent is a web in which anything that is the product of the will is surrounded and held up and partly formed by things that are not, in such a way that reflection can only go in one of two directions: either in the direction of saying that responsible agency is a fairly superficial concept, which has a limited use in harmonizing what happens, or that it is not a superficial concept, but that it cannot ultimately be purified—if one attaches importance to the sense of what one has done and what in the world one is responsible for, one must accept much of what makes its claim on that sense solely in virtue of being actual.[31]

31. "Moral Luck," pp. 29-30.

CAST OF CHARACTERS
(in order of appearance)

Sartre's student, *an occupied Parisian*

Lalitha, *a conservationist*

Keith Richards, *a rock star*

Paul Gauguin, *a painter*

Catherine Earnshaw, *a young gentlewoman*

Ellen Dean, *her governess*

Edgar Linton, *her fiancé*

Heathcliff, *her lover*

Minerva McGonagall, *a professor of Transfiguration*

Neville and Hermione, *her students*

Alexei Karenin, *a civil servant*

Anna Karenina, *his wife*

Joe, *a fair-weather friend*

Margaret Wilcox (*née* Schlegel), *a young lady*

Henry Wilcox, *her husband*

Billy and Suzy, *wayward children*

Black, *a wicked neuroscientist*

Martin Luther, *a priest*

BIBLIOGRAPHY

Bok, Hilary. "Freedom and Practical Reasoning." In Watson, ed. *Free Will*. Oxford, 2003.

Bratman, Michael. *Structures of Agency*. Oxford, 2007.

— "Autonomy and Hierarchy." In his *Structures of Agency*.

Brewer, Talbot. *The Retrieval of Ethics*. Oxford, 2009.

Broome, John. "Reasons." In Wallace, Scheffler, and Smith, eds. *Reasons and Value: Themes from the Moral Philosophy of Joseph Raz*. Oxford, 2004.

Campbell, C. A. *On Selfhood and Godhood*. George Allen and Unwin, 1957.

Chang, Ruth. "Voluntarist Reasons and the Sources of Normativity." In Sobel and Wall, eds., *Reasons for Action*. Cambridge 2009.

Chisholm, Roderick. "Human Freedom and the Self." In Watson, ed. *Free Will*. Oxford, 2003.

Darwall, Stephen. *The Second-Person Standpoint*. Harvard, 2006.

Ebels-Duggan, Kyla. "Against Beneficence: A Normative Account of Love." *Ethics* 119 (2008).

Fine, Kit. "Vagueness, Truth, and Logic." *Synthese* 30 (1975).

Fischer, John Martin, and Ravizza, Mark. *Responsibility and Control: A Theory of Moral Responsibility*." Cambridge, 1998.

— "The Frankfurt Cases: The Moral of the Stories." *The Philosophical Review* 119 (2010).

Frankfurt, Harry. "Identification and Wholeheartedness." In his *The Importance of What We Care About*. Cambridge, 1988.

— "On Caring." In his *Necessity, Volition, and Love*. Cambridge, 1999.

— "The Faintest Passion." In *Necessity, Volition, and Love*.

— *Taking Ourselves Seriously and Getting it Right*. Stanford, 2006.

Franzen, Jonathan. *Freedom*. Macmillan, 2010.

Frye, Marilyn. "A Note on Anger." In her *The Politics of Reality: Essays in Feminist Theory*. Crossing, 1983.

Hieronymi, Pamela. "The Force and Fairness of Blame." *Philosophical Perspectives* 18 (2004).

— "Controlling Attitudes." *Pacific Philosophical Quarterly* 87 (2006).

Hume, David, eds. Norton, D., and Norton, M. *A Treatise of Human Nature*. Oxford, 2000.

Jollimore, Troy. *Love's Vision*. Princeton, 2011.

Kane, Robert. *The Significance of Free Will*. Oxford, 1996.

— "Responsibility, Luck, and Chance: Reflections on Free Will and Determinism." *The Journal of Philosophy* 96 (1999).

Kolodny, Niko. "Love as Valuing a Relationship." *The Philosophical Review* 112 (2003).

Korsgaard, Christine. *The Sources of Normativity*. Cambridge, 1996.

— *Self-Constitution*. Oxford, 2008.

Kripke, Saul. *Wittgenstein on Rules and Private Language*. Harvard, 1982.

Langton, Rae. "Love and Solipsism." In Lamb., ed., *Love Analyzed*. Westview, 1997.

Lewis, C. S. *The Four Loves*. Harcourt Brace, 1960.

Lewis, David. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volumes* 63 (1989).

— "Are We Free to Break the Laws?" In Watson, ed. *Free Will*. Oxford, 2003.

McDowell, John. "Non-cognitivism and Rule-Following." In his *Mind, Value, and Reality*. Cambridge, 1998.

Milgram, Elijah. *Practical Induction*. Harvard, 1997.

Nagel, Thomas. *The View from Nowhere*. Oxford, 1986.

Nehamas, Alexander. "The Good of Friendship." *Proceedings of the Aristotelian Society* 90:3 (2010).

Nietzsche, Friedrich, trans. Faber, M. *Beyond Good and Evil*. Oxford, 1998.

Nozick, Robert. *Philosophical Explanations*. Harvard, 1981.

Nussbaum, Martha. "Love and the Individual: Romantic Rightness and Platonic Aspiration." In her *Love's Knowledge*. Oxford, 1990.

Pereboom, Derk. "Determinism *al Dente*." *Noûs* 29 (1995).

— *Living Without Free Will*. Cambridge, 2001.

Rawls, John. *A Theory of Justice* (revised edition). Harvard, 1999.

Raz, Joseph. *Engaging Reason*. Oxford, 1999.

— "Incommensurability and Agency." In his *Engaging Reason*.

Richards, Keith. *Life*. Little, Brown, and Company, 2010.

Richardson, Henry. "Specifying Norms as a Way to Resolve Concrete Ethical Problems." *Philosophy and Public Affairs* 19 (1990).

Rorty, Amelie Oksenberg, "The Historicity of Psychological Attitudes: Love is Not Love Which Alters Not When It Alteration Finds." *Midwest Studies in Philosophy* 10 (1987).

Sartre, Jean-Paul, trans. Macomber, C. *Existentialism is a Humanism*. Yale, 2007.

Scanlon, T. M. "The Significance of Choice." In Watson, ed. *Free Will*. Oxford, 2003.

— *Moral Dimensions*. Harvard, 2008.

Smith, Angela. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115 (2005).

Smith, Michael, and Pettit, Philip. "Freedom in Belief and Desire." In Watson, ed. *Free Will*. Oxford, 2003.

Strawson, Galen. "The Impossibility of Moral Responsibility." In Watson, ed. *Free Will*. Oxford, 2003.

Strawson, P. F. "Freedom and Resentment." In his *Freedom and Resentment and Other Essays*. Methuen, 1974.

Taylor, Charles. "Responsibility for Self." In Rorty, ed., *The Identities of Persons*. Berkeley, 1976.

Van Inwagen, Peter. "An Argument for Incompatibilism." In Watson, ed. *Free Will*. Oxford, 2003.

Velleman, J. David. "Love as a Moral Emotion." *Ethics* 109 (1999).

— "Beyond Price." *Ethics* 118 (2008).

Walker, Margaret Urban. "Resentment and Assurance." In Calhoun, ed., *Setting the Moral Compass: Essays by Women Philosophers*. Oxford, 2004.

Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Harvard, 1994.

— "Dispassionate Opprobrium: On Blame and the Reactive Sentiments." In Wallace, Kumar, and Freeman, eds., *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*. Oxford, 2011.

Watson, Gary, "Responsibility and the Limits of Evil." In Schoeman, ed., *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge, 1988.

— "Free Agency." In Watson, ed. *Free Will*. Oxford, 2003.

— "Introduction." In Watson, ed. *Free Will*. Oxford, 2003.

Wiggins, David. "Truth, Invention, and the Meaning of Life." In his *Needs, Values, Truth: Essays in the Philosophy of Value* (third edition, amended). Oxford, 2003.

Williams, Bernard. "The Makropulos Case: Reflections on the Tedium of Immortality." In his *Problems of the Self*. Cambridge, 1973.

— "Persons, Character, and Morality." In his *Moral Luck*. Cambridge, 1981.

— "Moral Luck." In *Moral Luck*.

— "Internal and External Reasons." In *Moral Luck*.

— "How Free Does the Will Need to Be?" In his *Making Sense of Humanity*. Cambridge, 1995.

— "Nietzsche's Minimalist Moral Psychology." In *Making Sense of Humanity*.

— "Internal Reasons and the Obscurity of Blame." In *Making Sense of Humanity*.

Wolf, Susan. "Sanity and the Metaphysics of Responsibility." In Watson, ed. *Free Will*. Oxford, 2003.

— "Blame, Italian Style." In Wallace, Kumar, and Freeman, eds., *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*. Oxford, 2011.