Self to Self

Selected Essays

J. DAVID VELLEMAN
New York University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521670241

First published 2006
Reprinted 2007

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data

Velleman, James David.
Self to Self : selected essays / J. David Velleman.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-85429-6 (hardcover) – ISBN 0-521-67024-1 (pbk.)
1. Self. 2. Self (Philosophy) 3. Kant, Immanuel, 1724–1804 – Ethics. I. Title.
BF697.V45 2005
126–dc22 2005008114

ISBN 978-0-521-85429-0 Hardback
ISBN 978-0-521-67024-1 Paperback

Transferred to digital printing 2009

11

The Centered Self

> In that demand he was obeying the voice of his rigid conscience, which had never left him perfectly at rest under his one act of deception – the concealment from Esther that he was not her natural father, the assertion of a false claim upon her. 'Let my path be henceforth simple,' he had said to himself in the anguish of that night; 'let me seek to know what is, and if possible to declare it.'
>
> – George Eliot, *Felix Holt*

We have many expressions to describe a person who is trustworthy and true – a *rock,* a *brick,* a *Mensch.* In a more analytical mood, we describe such a person as *grounded* or *centered.* I want to consider what it is to

An ancestor of this chapter, entitled "A Sense of Self," was presented as one of the Jerome Simon lectures at the University of Toronto; to a conference on personal identity and practical reason at the University of Illinois, Chicago; to the Moral Philosophy Seminar at Oxford University; and to the philosophy departments at the University of Virginia, NYU, and Tufts University. "A Sense of Self" was the target of a paper delivered by Maik Tåndler to the Göttinger Philosophisches Kolloquium in January 2003, where much helpful discussion ensued; and it was the topic of discussion at a September 2003 meeting of the Ohio Reading Group in Ethics. Thanks are due to Ted Hinchman, Jim Joyce, Dick Moran, and Thomas Schmidt for extensive comments on drafts of that essay.

The present chapter was delivered at the University of Michigan; to the philosophy departments of the University of Saskatchewan, the University of California at Riverside, the University of Dundee, the University of Stirling; the University of Edinburgh, the University of St. Andrews, and the University of Bristol; at a conference on Values, Rational Choice, and the Will at the University of Wisconsin, Stevens Point; and at the 2004 Oberlin Colloquium, where the commentator was Tom Hill. This chapter contains material from the "Precis" and "Replies" that I contributed to a symposium on my book *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000). The symposium, with commentaries by Jonathan Dancy, Alfred Mele, and Nadeem Hussain, was published in *Philosophical Studies* 121 (2004).

be grounded or centered, and then to explain what being grounded or centered has to do with being trustworthy and true.

My account begins with a quality generally regarded as distinctive of persons – namely, self-awareness.[1] Of course, a brick or a rock isn't self-aware; but a person can be a brick or a rock in the figurative sense only through the utmost development of that which differentiates him as a person from bricks and rocks literally so called. If we want to identify the relevant differences, however, we do better to contrast a person with something that comes a bit closer to personhood – say, a cat.

Now, a cat is conscious, I assume, and it has the sort of consciousness whose content can be put into words only with the help of the first-person pronoun. A cat could never catch a mouse if it couldn't have thoughts representing the world from its own egocentric perspective, thoughts with English-language equivalents such as "I'm gaining on it" or "I've got it." There is a sense, then, in which a cat has first-personal awareness. A cat can even have a reflexive awareness of a sort, as when it realizes that the tail it has been chasing is its own.

[1] This section is heavily indebted to Thomas Nagel's work on "the objective self" and John Perry's work on self-knowledge. I include a discussion of Perry in Appendix A. In the remainder of this note I'll briefly summarize my debt to Nagel.

Nagel has argued that the self is that part or aspect of a person that harbors his objective conception of the world. This conception provides the mental context for the question "Who am I?" When a person asks himself "Who am I?" he is in effect asking "Which person am I?" while surveying the possible candidates from an impartial distance. "Who am I?" must therefore be understood as spoken from a standpoint that's objective in the sense that it views all persons from the outside as possible referents for the pronoun 'who'. And the 'I' in this question must emanate from that part or aspect of a person which occupies this stance, surveying people from a distance and seeking to identify with one of them.

This conception of oneself, as a person among others, figured in Nagel's first book, *The Possibility of Altruism*, as the starting point of moral thought. There, Nagel argued that the conception of oneself as a person among others constrains one's practical reasoning in the manner of Kant's Categorical Imperative. If this argument is combined with the premise of Nagel's argument about the self, the result is a conclusion about the source of morality. The conclusion is that moral constraints on practical reasoning are imposed by nothing other than one's sense of identity. My aim in this chapter can be described in the same terms.

See Nagel, "Subjective and Objective," in *Mortal Questions* (Cambridge: Cambridge University Press, 1979), 196–213; 'The Limits of Objectivity," in *The Tanner Lectures on Human Values*, Vol. I, ed. S. McMurrin (Salt Lake City: University of Utah Press, 1980), 77–139; "The Objective Self," in *Knowledge and Mind*, ed. Carl Ginet and Sydney Shoemaker (New York: Oxford University Press, 1983), 211–32; *The View From Nowhere* (New York: Oxford University Press, 1986), Chapter IV. *The View From Nowhere* is perhaps the most widely read of these works, but its chapter on the "objective self" is, in my view, considerably watered down. I recommend the essay entitled "The Objective Self" in the volume edited by Ginet and Shoemaker.

What a cat lacks, however, is a conception of a creature that it is. A cat is aware of the mouse that it is chasing, but it is not aware of there being a creature by whom the mouse is hereby being chased. When a cat recognizes its own tail, it merely forges a mental association between an object seen to its rear and a locus of sensation or motion at its rear end. It has no conception of being a creature chasing its own tail.

By contrast, when a person realizes that he's stepping on his own shoelaces, he attains more than a mental association between the sensation of treading on something with one foot and the sensation of being tripped up in the other. He has the concept of a particular person bearing the name to which he answers, sporting the face that looks back at him from the mirror, and doing the things that he is aware of doing – including, at the moment, stepping on his own shoelaces. Unlike a cat, a person is aware of being somebody, and he usually knows a fair amount about the somebody who he is.

A person's conception of who he is constitutes the axis on which he can potentially be centered, or the anchor by which he can potentially be grounded. Here I hope to be saying nothing new. I take it to be part of the ordinary concept of being grounded or centered that these qualities depend on a person's sense of identity. Less obvious, perhaps, is that a person's sense of identity involves an objective conception of someone in the world who he is – a particular, persisting member of the objective order to whom he can pin the unseen point at the center of his point-of-view. What is not at all obvious, and what I hope to explain, is how pinning his point-of-view to that person can make him a rock or a brick or a *Mensch*, trustworthy and true.

In order to explore this question, I'll need an example of a situation that (you should pardon the expression) separates the *Menschen* from the boys and girls. I'm going to use the most familiar example that I know of – the prisoners' dilemma. My goal is to show how our understanding of this tired example can be refreshed by reflection on the nature of human self-awareness. I'll start with a quick review of how the dilemma comes about.

Suppose that you and I find ourselves in circumstances where each would lose something by cooperating with the other, no matter what the other does, but would lose even more from the other's failure to cooperate. The cooperation at issue might be helping to harvest one another's fields or, to invoke the relevant cliché, merely scratching one another's backs. In these circumstances, neither of us has anything to

gain from helping the other, whether or not the other helps us, and both of us therefore face the prospect of the other's refusing to help. We might wish that we could escape the dilemma through an exchange of mutually dependent offers of the form "I will cooperate if you will."[2] As is well known, however, the resulting agreement would generate a second-order dilemma, since each of us would lose by following through on the agreement, though he would lose even more from the other's refusal to follow through.

Assume that none of the usual devices for resolving our dilemma is available – no past experience with one another, no external sanctions against cheating, no future opportunities for retaliation or repayment. Assume, in other words, that ours is a classic, one-time prisoners' dilemma, in which the parties have knowledge of nothing but the payoffs and one another's rationality. The point of this assumption, for my purposes, is to deprive us of any social, emotional, or indeed moral resources for coping with our dilemma, not because such resources are absent from dilemmas in real life but because their absence from this imagined dilemma will force us to rely on resources of the solitary, even solipsistic kind to which centeredness and groundedness belong. My exclusive

[2] I will discuss a version of the dilemma in which the parties are given the opportunity to make a cooperative agreement, if they can; and my resolution of the dilemma will ultimately depend on the rationality of making and then abiding by such an agreement. Hence my discussion of the prisoners' dilemma is not about the rationality of cooperation *per se*; it's about the rationality of truth-telling and constancy in agreements. I do not try to show that acting cooperatively is rational in itself; I try to show only that it can be made rational by the exchange of commitments that are in turn rational for the parties to exchange and then to carry out.

This distinction is essential to coordinating the present discussion with the discussions of Kantian ethics elsewhere in this volume. As I explain in "A Brief Introduction to Kantian Ethics" (Chapter 2 in the present volume), a moral requirement to cooperate in the prisoners' dilemma must be derived, in Kantian theory, from a contradiction in the will. A universal law of non-cooperation is not impossible in itself, and so Kantianism must find a rational obstacle to our willing there to be such a law. (See also "Willing the Law" [Chapter 12 in the present volume].) But as I also explain in the "Brief Introduction," moral strictures against breaking commitments and lying are derived from contradictions in conception – that is, from the impossibility of there being universal laws for these practices rather than from our inability to will such laws. Since my resolution of the prisoners' dilemma in this chapter depends on the rationality of truth-telling and constancy in cooperative agreements, rather than the rationality of cooperative action in itself, my argument will correspond to the Kantian derivation of a contradiction in conception rather than a contradiction in the will. My remarks in the "Brief Introduction" to the effect that prisoners' dilemmas generate contradictions in the will are about the morality of acting cooperatively in such dilemmas, not the morality of making and keeping agreements to do so.

focus on these resources should not be taken to imply that they are the only resources available for coping with prisoners' dilemmas.

The idea of offering to cooperate in these circumstances is not entirely daft. Mutually beneficial cooperation would be possible if only we had, and knew that we had, two crucial abilities. First, each of us would need the ability to form an effective conditional intention, to cooperate if the other formed a reciprocal and equally effective intention.[3] By "an effective intention" I mean an intention that would determine the course of the subject's future behavior – in this case, by determining the subject to cooperate if he knew that the other party intended likewise. Second, each would need the ability to let the other know his state of mind. By "to let the other know his state of mind" I mean making his state of mind evident to the other so as to instill in him a true and reliably justified belief as to whether the condition on his own intention had been fulfilled. If we had these two abilities, and our having them was common knowledge between us, then each of us would have good reason to form the conditional intention to cooperate if the other intended likewise, and then to let the other know of that intention, by saying "I'll cooperate if you will."[4] Each party's intention would lead him to bear the cost of actually cooperating only if its condition were fulfilled by the other's intention, in which case it would fulfill the condition of the other's intention, thereby leading to the greater benefit of the other's cooperation. The costs of committing himself to cooperate would therefore be appropriately linked to overriding benefits, which would accrue from triggering the other's commitment.

This calculation is what gives rise to the idea of saying "I'll cooperate if you will." Unfortunately, the calculation reckons on our having abilities that can seem impossible for us to have. How can I determine my future behavior by means of a present intention? And how can I give you reliable grounds for believing that I have such an intention? In any cooperative agreement, the benefit to me flows from your believing in my effectively

[3] Note that each commitment is conditional on the other speaker's commitment rather than his action. That is, each says "I will cooperate if you will," not "... if you do." Hence the condition on each commitment is satisfied as soon as the other commitment is issued. I discuss such commitments at length in the Appendix to "Deciding How to Decide," reprinted in *The Possibility of Practical Reason*, 242–43.

[4] Here I am assuming that, although we have the ability to make our intentions known to one another, we do *not* have the ability to lead one another to believe in intentions that we do not actually have. The latter ability would enable us to skip the step of forming a cooperative intention before expressing it.

intending to cooperate, not from my actually intending to cooperate, and certainly not from my so intending effectively. Even if I formed an intention to cooperate, I would have no reason to let it take effect in my future behavior, and I have no reason to form a cooperative intention if I can convincingly feign one instead. It therefore seems that I cannot commit my future self to cooperate, and that, even if I could, I cannot give credible evidence of having done so. A classic, one-time prisoners' dilemma thus generates two problems – a problem of commitment and a problem of credibility – neither of which appears to be soluble in the circumstances.

What makes these problems seem insoluble, however, is the instrumental conception of practical reasoning as a calculation of costs and benefits, a conception that narrows the range of considerations available to us as participants in the dilemma. We are in fact capable of making rationally effective commitments and of giving one another rational grounds for believing in them. Not surprisingly, our capacity to be credibly committed depends on our capacity to be centered or grounded, which in turn depends on the sense of identity made available to us by our distinctively human form of self-awareness. The problem with the instrumental conception of practical reasoning is that it affords no role for our sense of identity to play, and hence no role for our capacity to be centered or grounded. No wonder, then, that it makes credible commitments seem impossible. What's needed is a conception of practical reasoning that has a role for our sense of identity, which might in turn explain our capacity for credible commitments. So let's examine the connection between practical reasoning and self-awareness.

As we have seen, self-awareness gives me an objective conception of the person who I am. That conception bears on practical reasoning, to begin with, by giving me access to objective knowledge of what I am doing.

Of course, a cat is also aware of doing things, such as hissing at someone by whom it feels threatened. But a cat's awareness of its own doings never extends to the knowledge that they are being done by a creature in the world. It represents them from the perspective of the one doing them, without representing the creature occupying that perspective. Thus, even when a cat is aware of hissing at you, and even if it is hissing with the thought of scaring you away, it cannot be thinking that you will be scared of this hissing creature – scared, that is, of its hissing self – because it has no conception of being one of the world's creatures, and hence no sense of self. By contrast, if I tried to scare you away, I would be aware of

confronting you with a person saying "Scram!" as would be manifest in that very utterance, since a person saying "Scram!" is intimidating precisely by virtue of manifesting the intention to be an intimidating person.

In performing a communicative action of this kind, I must be able to understand what I am doing as I intend it to be understood by you. In order to tell whether my behavior might be understood as my trying to scare you away, I must find it potentially understandable in those terms, vicariously sharing the understanding that I intend to elicit. This shared understanding requires me to conceive of what I'm doing as done by the creature who I am, a creature who might potentially scare you away by saying "Scram!" – which is different from conceiving merely of doing it, from the perspective of the unrepresented do-er.

Along with the ability to understand what I'm doing as done by the creature who I am comes the possibility of finding it unintelligible in those terms. A cat can round on its own tail and wonder, "What is that thing up to?" But I can round on my entire self and wonder, "What is this creature up to?" As soon as a cat associates the waving motion that it sees to its rear with the motion that it is aware of making from its rear end, its puzzlement is over. It knows why the tail is waving, since it is now aware of waving it. It cannot go on with "Yes, but why am I waving my tail?" That question would be about the behavior of a tail-waving creature, which it has no cognizance of being. Self-puzzlement of this latter kind is possible only for a creature whose awareness of doing things results in an awareness of their being done by the creature who he is.

I think that the state of mind variously described as puzzlement, mystification, confusion, perplexity, or bewilderment deserves more philosophical attention than it ordinarily receives. This state is aversive: we try to avoid it, and when we have gotten into it, we try to get out. The aversiveness of this state is a reminder that we have intellectual drives. We do not passively receive knowledge; we gain it through cognitive activity, driven by intellectual impulses. And the frustration of these impulses is aversive, like the frustration of any fundamental drive.

A human being's intellectual impulses are sometimes directed at the person who he is. The creature with whom he is aware of being identical naturally has a special salience for him – as the creature walking in his shoes, sleeping in his bed, eating his meals – and the doings of that creature therefore become the object of his intellectual drives. But the person's awareness of being identical with that creature opens up an obvious shortcut to knowledge about its doings. He must realize that

doing things – that is, behaviors conceived from his perspective as the unrepresented agent – constitutes their being done by that creature, the same behavior conceived objectively. And he must realize that seeking to know what it is doing – an intellectual activity conceived from his perspective as the unrepresented inquirer – constitutes that creature's striving for self-knowledge. Finally, then, he must realize that he can know what that creature is doing simply by doing what he conceives of it as doing, or as being about to do, since his conception will then turn out to be not only true but also justified, on the grounds of the creature's having this very intellectual incentive to bear it out. He tends to behave as he conceives of that creature as behaving because he will then have, embodied in that conception, a knowledge of what that creature is doing; and that conception will have the reliability of knowledge because it is about a creature for whom the prospect of having knowledge embodied in it is an incentive to behave accordingly.

Strange as this psychological mechanism may sound, it has been copiously documented by social psychologists working in the area that is sometimes labeled "self-consistency," an area encompassing the topics of cognitive dissonance and attribution. Research in this area has shown that people have a broad tendency to behave in ways that cohere with their own conceptions of themselves – of how they behave in general and of their motives on a particular occasion. Potential voters are more likely to vote in an election if they have antecedently predicted that they are going to. Children are more likely to be tidy if told that they *are* tidy than if told that they ought to be. People behave angrily if they are led to believe that they are angry – the more angrily, the more angry they are led to believe they are. Shy people don't behave shyly if they are led to attribute the symptoms of their social anxiety to other causes. And so on.[5]

One team of researchers has observed that subjects' behavior can be influenced by the act-descriptions that they are antecedently prompted to frame, as if they have a tendency to fulfill antecedently framed descriptions of their forthcoming actions.[6] This tendency is cited by the researchers to explain how people know what they are doing – which is the very explanation that I have just offered: people know what they're doing because they tend to do what they have just now thought that they

[5] I discuss these and other empirical results in "From Self Psychology to Moral Philosophy" (Chapter 10 in the present volume).

[6] See the publications of Wegner, Vallacher, and colleagues cited at notes 59 and 60 of "From Self Psychology to Moral Philosophy" (Chapter 10 of the present volume).

are just about to do. The psychologists give this mechanism the label "act identification." And they invoke this mechanism, not only to explain how people generally know what they are doing, but also as a model for the process of acting on an intention: to frame an act-description and then fulfill it, they suggest, is just to form an intention and act on it.

With this reference to acting on an intention, we begin to see the true relevance of self-awareness to practical reasoning. Because I have an objective conception of the creature who I am, I can be puzzled by the behavior of that creature, but I can also avoid such puzzlement by first framing an idea of the creature's next action and then enacting that idea, a process that social psychologists have observed and have identified with the process of forming and acting on an intention. And acting on an intention is the consummation of practical reasoning.

This model of intention illustrates a central thesis of the book entitled *Intention,* by Elizabeth Anscombe. In that book, Anscombe analyzes the difference between what we do and what merely happens to us, or in us. The difference, she argues, is that our doings are the object of a special kind of knowledge, which Anscombe calls "knowledge without observation."

Anscombe uses the notion of knowledge without observation to explain the difference between two kinds of indicative statements about the future: expressions of belief, such as "I'm going to be sick," and expressions of intention, such as "I am going to take a walk" (p. 1). If someone responds to the statement "I am going to be sick" by asking "Why would you do a thing like that?" he has misinterpreted the speech act, by failing to recognize it as an expression of belief rather than intention. Conversely, if someone responds to "I am going to take a walk" with "How can you tell?" he has failed to recognize it as an expression of intention rather than belief. Now, the difference between these statements cannot lie in the former's being informative and hence potentially knowledge-conveying, since the latter is also informative and hence potentially knowledge-conveying. As Anscombe puts it, "the indicative (descriptive, informatory) character is not the distinctive mark of 'predictions' *as opposed to* 'expressions of intention', as we might at first sight have been tempted to think" (§2, p. 3).

In Anscombe's view, the difference between "I am going to take a walk" and "I am going to be sick," given that both can convey knowledge possessed by the speaker, is that the knowledge conveyed by the latter is speculative, whereas the knowledge conveyed by the former is practical, in

the sense that it causes the facts that make it true (§48, p. 87). "I am going to be sick" expresses a belief that is caused by evidence of the speaker's becoming sick, whereas "I am going to take a walk" expresses an intention that causes the speaker to take a walk. In expressing this intention, however, the speaker is also expressing his knowledge of what he is going to do, which must therefore be "known by the being the content of [his] intention" (§30, p. 53). Hence the speaker has knowledge embodied in a mental state that causes – rather than being caused by, or causally concomitant to – the facts that make it true (§48, p. 87).[7] Knowledge that is thus productive rather than receptive of what is known is what Anscombe has in mind when speaking of "knowledge without observation."

Why might one be tempted to think of agency in this way? Anscombe attributes her use of the phrase "practical knowledge" to Aquinas, for whom the phrase described God's knowledge of His creation. God knows what the world is like, but not by dint of having found out; He knows what the world is like because it is just as He means it to be. And His meaning it to be that way already constitutes knowledge on His part of how it is. This epistemological relation that God bears to the world – knowing how it is just by meaning it to be that way – is constitutive of His role as the world's designer. The designer of something is the one whose conception of the thing determines how it is, rather than vice versa, and determines this by a mechanism reliable enough to justify his confidence in that conception as an accurate representation. To be the designer of something is just to be the one whose conception of it has epistemic authority by virtue of being its cause rather than its concomitant or effect.

Anscombe's nod to medieval theology as her source for the term "practical knowledge" suggests that she conceives of intentional action as a realm in which human beings exercise a minor share of divinity. We create our intentional actions, just as God creates the world, and our creating them consists in our framing a conception of them that has epistemic authority by virtue of being determinative of them.

What I have sometimes presumed to call my theory of agency is little more than a variation on this theme of Anscombe's. My main departure from Anscombe has been to introduce a story about the dynamics of practical

[7] It is important not to confuse practical knowledge, in Anscombe's sense of the term, with practical wisdom, or *phronesis*, as discussed by Aristotle. Practical knowledge, also called maker's knowledge, is distinguished not by its subject matter but by its causal relation to its object. When judged by this causal relation, Aristotelian practical wisdom is actually theoretical rather than practical, since it is receptive rather than productive of the facts known.

knowledge – the story that I have just now been telling, of how our actions are guided by our conceptions of them because of our intellectual drives toward the knowledge that is consequently embodied therein.

The same researchers who claim to have observed this process in action also claim to have shown that we ordinarily seek to identify our behavior at a "high" or "comprehensive" level, representing our underlying motives and ultimate goals. They describe this further tendency as a "search for meaning in action"[8] or "a human inclination to be informed of what we are doing in the most integrative and general way available."[9] Here the empirical findings harmonize with my dynamic version of Anscombe's theory in a further respect.

With a now famous example, Anscombe points out that an agent often knows what he is doing under a series of descriptions each of which incorporates the answer to the question "Why?" directed at the same action under the previous description in the series. Why is he moving his arm? Because he is pumping water. Why is he pumping water? Because he is replenishing the water supply. Why is he replenishing the water supply? Because he is poisoning the inhabitants of the building. Why is he poisoning the inhabitants? Because he is assassinating enemy agents. And so on. With the exception of the first, purely physical description, all of the descriptions under which this person knows what he's doing are answers to the question why he is doing it as previously described.[10]

The sequence from "moving his arm" to "killing enemy agents" displays a progression toward increasingly "high-level" or "comprehensive" act-descriptions. So if there is empirical evidence of "a human inclination to be informed of what we are doing in the most integrative and general way available," as the act-identification theorists claim, then it is evidence of an inclination to progress from rudimentary descriptions like the former toward comprehensive descriptions like the latter.

I believe that the existence of such an inclination follows directly from our having intellectual impulses directed at the behavior of the person

[8] Wegner & Vallacher, "Action Identification," in *Handbook of Motivation and Cognition*, ed. Richard M. Sorrentino and E. Tory Higgins (New York: Guilford Press, 1986), pp. 555–56.

[9] Vallacher & Wegner, *The Theory of Action Identification* (Hillsdale, NJ: Erlbaum, 1985), p. 26.

[10] Mere act descriptions do not amount to explanations, of course. The successive descriptions in Anscombe's example are descriptions under which the agent's action is intentional, and it is the corresponding intentions on his part that explain his action. When the act-descriptions are spoken in the first person – "I am pumping water," and so on – they express the relevant intentions, but a complete explanation would have to cite those intentions rather than merely express them.

who we are. The object of our intellectual drives must be, not merely the recording of rudimentary, observable facts, but also the development of "integrative and general" ways of formulating them. When directed at our own behavior, these drives must demand a knowledge of what we are doing in the sort of comprehensive terms that also explain why we are doing it. And the previously described shortcut to self-knowledge – the shortcut of doing what we think we are doing, or are about to do – is also a route to this "high level" self-knowledge. For we can attain integrative knowledge of what we are doing simply by framing and fulfilling integrative conceptions of our behavior, conceptions formulated in terms of the dispositions and circumstances that help to explain it.

In order to frame and fulfill integrative conceptions of our behavior, of course, we must be aware of relevant factors with which to integrate it – desires by which it might be motivated, emotions that it might express, customs and policies that it might implement, traits of character that it might manifest. These other aspects of our self-conception – motives, emotions, customs, policies, traits of character – can fill out an integrative knowledge of what we are doing, provided that we do things appropriately integrated with them. The drive toward a more comprehensive knowledge of what we are doing therefore favors doing things that can be understood as motivated by our desires, expressing our emotions, implementing our policies, manifesting our characters, and so on.

Aspects of ourselves and our circumstances that could fill out an integrative conception of doing something turn out to coincide with what we ordinarily count as reasons for doing it. Examples of desire-based reasons are well known, but reasons can also be based on other considerations that would help to explain an action, as illustrated by these examples:

Why are you whistling?
Because I'm happy.

Why aren't you having any wine?
Because I don't drink.

Why worry about his problems?
Because I'm his friend.

Why are you shaking your head?
Because I think you're wrong.

Why do you have her picture on your wall?
Because I admire her.

Here already?
I'm punctual.

I believe that reasons for doing something are facts that would inform an integrative knowledge of what we were doing, if we did that thing. Our intellectual drives favor framing and fulfilling a conception of ourselves as doing that thing, understood in the light of those facts, rather than other things for which we lack an equally integrative conception. Reasons for doing something are facts in light of which doing it would make sense.[11]

This concludes my account of how human self-awareness structures practical reasoning. Being driven to know what I am doing, and to know it in terms that explain why, I frame an explanatory conception of doing something and then I do it. My antecedent conception of doing something is my intention to act, and the explanatory facts on which it draws are my reasons for the intended action.

I now want to argue that this account provides the resources for resolving or at least mitigating the prisoners' dilemma. Specifically, this account of practical reason provides resources for attacking the problems of commitment and credibility, which stand in the way of our reaching a cooperative agreement. I will begin with the problem of credibility, assuming for the moment that the problem of commitment can be solved; I will then turn to the latter problem.

According to the traditional understanding of the prisoners' dilemma, neither of us has any reason to take the first step of saying "I will cooperate if you will." If I made this offer, you would know that I stood to lose by following through, and so you would suspect that rationality would lead me to default. Indeed, you would know that I must already intend or at least expect to default, since the costs of following through must be as obvious to me as they are to you. My offer would thus be transparently insincere, and so it would elicit nothing from you in return, except perhaps an offer of equally transparent insincerity. The whole exchange would therefore be pointless, as would be common knowledge between us.

But suppose that I nevertheless proceeded to say "I will cooperate if you will." How might you understand my utterance?

You might consider the possibility that I was doing something superficially pointless, by offering to cooperate, for the deeper purpose of signaling a genuine intention to do something outright irrational, by following through on that offer. But any thought you might entertain of attributing cooperative intent to my utterance is a thought that I should have foreseen and thought of exploiting to my advantage. The thought

[11] Since the main purpose of this chapter is to apply this conception of reasons for acting, not to defend it, I have relegated objections and replies to Appendix B.

of attributing cooperative intent to my utterance would therefore lead you to the opposite hypothesis, that I hoped to elicit that very attribution in order to take advantage of you – a train of thought that I should have foreseen, thereby foreseeing that my utterance would be fundamentally pointless, after all.

Knowing that my offer was pointless, you might well ask yourself, "What on earth is he doing?" But you would also know that the pointlessness of my offer was known to me – knowledge that should have left me, as it left you, at a loss to understand what I was doing. The question that you pose to yourself might therefore be not just "What on earth is he doing?" but "What on earth does he *think* he is doing?"

Now, there is a significant difference between the questions "What are you doing?" and "What do you think you are doing?" The former is a straightforward request for information, but the latter is often an expression of protest or surprise. This question expresses protest or surprise because a rational agent is normally expected to do things that he can understand. If someone's action makes no sense to us, we are prepared to believe that the failure is ours and can be remedied by more information, which is usually available from him; but if we cannot see how his action could make sense to him, then we believe that the failure is his, and that it is a failure not merely of intellect but of action, a failure not just to understand what he's doing but also to do what he can understand. We are surprised to find him doing something that he himself cannot understand, and our asking "What do you think you are doing?" expresses our surprise at his doing it.

According to the theory outlined here, having an answer to this question is the cognitive goal to which there is an irresistible shortcut that is constitutive of practical reasoning. Hence the assumption that I should be able to answer the question follows from the assumption that I am rational in a sense derived from the foregoing account of practical reasoning. And as parties to a classic prisoners' dilemma, you and I are allowed to assume one another's rationality. These mutual assumptions of rationality can now be reinterpreted, as assumptions of one another's tendency to act so as to understand what he's doing, by doing what he has the resources to understand. When our mutual assumptions of rationality are reinterpreted in this way, our dilemma takes on a new complexion.

Thus far, the discussion of whether a cooperative offer would be intelligible has proceeded on a familiar assumption about how behavior can be understood. The assumption has been that in order to understand what

I'm doing, in offering to cooperate, you and I must find desired consequences to which I might regard the offer as instrumental. Since I can't expect my offer of cooperation to be taken seriously, I can't regard it as instrumental to anything I want, and so I have seemed unequipped to understand it. Yet the conception of practical reasoning as the shortcut to self-understanding does not presuppose that behavior must be understood instrumentally; it can accommodate the fact that behavior is often understood in other ways.

For present purposes, the relevant alternative is to understand behavior expressively. For example, my belief that there's leftover chili in the fridge involves a motivational disposition to go to the fridge if I want some chili; but it also involves an expressive disposition to think or say "It's in the fridge" if a question arises about the availability of chili. The expressive disposition associated with belief is what causes us on occasion to say what we think even though we have no desire to communicate it – indeed, to blurt out what we think despite a positive desire to keep it to ourselves. As the latter case suggests, this expressive disposition is antecedent to any practical reasoning. This expressive disposition may actually conflict with the motivational disposition associated with the very same belief. For example, suppose that there's only one serving of chili left, and I want to eat it. If asked "Is there any chili left?" in that case, I can act instrumentally and say "No" while sidling toward the fridge, or I can act expressively and say "It's in the fridge." The disposition toward the latter, expressive behavior is what I might have to restrain by (as we say) biting my tongue in order to take the former, instrumental course. Either course of action will be intelligible, the one as motivated by my belief, the other as expressive of it.

If reasons for acting are considerations in light of which an action would make sense, then a belief can provide either instrumental or expressive reasons for acting, by rendering an action intelligible either as motivated by the belief or as expressive of it.[12] Asked whether there is any chili left, I may find expressive reason to say "It's in the fridge," if I believe there to be chili in the fridge, and I may find instrumental reason to say "No," if I also want the chili for myself. Which reason is stronger depends, in my view, on which action would allow for the best overall self-understanding.

In the case of our prisoner's dilemma, saying "I'll cooperate if you will" would make no sense when considered as motivated by desire and

[12] For a somewhat different view of expressive reasons for acting, see Robert Nozick, *The Nature of Rationality* (Princeton: Princeton University Press, 1993), 26–35.

belief, because it cannot rationally be expected to promote anything that I want; but it could easily be understood as the natural expression of an intention – specifically, an intention to cooperate if you should express a corresponding intention.[13] If I had such an intention, then my saying "I'll cooperate if you will" would be perfectly intelligible, as expressing my state of mind. The hypothesis that I have such an intention would therefore enable you to understand what I was doing. What's more, this hypothesis would enable you to understand how *I* could understand what I was doing; to understand how I could expect to be understood; and so on.

In short, it is common knowledge between us that my offer to cooperate would make sense if it expressed an intention to cooperate but not if conceived in purely instrumental terms, as a means to desired ends. And we have assumed it to be common knowledge that I am rational and therefore unlikely to do things that I don't understand. Since I could understand what I was doing, in offering to cooperate, only if I had the cooperative intention that my offer would express, you would have reason to assume that I had the intention and understood myself as expressing it. You would therefore have grounds for interpreting my offer as sincere, and I would have grounds for expecting it to be so interpreted. A solution to the problem of credibility appears to be at hand.

Unfortunately, this solution can be suspected of reviving the problem. For as soon as I have an expectation of being believed, I can have instrumental motives for offering to cooperate, and so I can understand making such an offer insincerely, without cooperative intent. If you might figure that I wouldn't offer to cooperate unless I had the intention that I could understand such an offer as expressing, then I can hope to gain the benefit of your cooperation by making the offer, and I can consequently understand it as motivated by that hope, even in the absence of any cooperative intention for it to express. Any nascent possibility of trust, or hope of being trusted, would thus appear to nip itself in the bud.

[13] Here I assume that the expressive disposition attached to beliefs is also attached to intentions. I base this assumption partly on my view that intention is a cognitive state that is similar to belief in taking its propositional content to be true, with the aim of so taking it only if it really is true. See my *The Possibility of Practical Reason*, esp. Chapters 1, 2, and 9. My conception of intention is borrowed from Anscombe, who bases it precisely on the observation that the natural way to express an intention is to assert that one is going to act.

Yet my revived instrumental understanding of an insincere offer would once again be unstable, precisely because its availability to me would be evident to you, as would in turn be evident to me, thus rendering the offer instrumentally pointless in my eyes. As soon as I begin to think instrumentally in this case, I enter a dizzying spiral of anticipating that my instrumental calculations have been anticipated, that their validity has thus been compromised, that their being so compromised has also been anticipated, with the result that they gain new validity, which has of course been anticipated, and so on. Hence the best instrumental understanding that I can achieve of what I am doing, if I offer to cooperate in these circumstances, is that I am taking a shot at being trusted, a shot whose prospects of success are obscured by endless complications. It is indeed a tangled web we weave, not only when we practice to deceive, but even when we practice honesty on instrumental grounds.

If I understand myself expressively, as intending to reciprocate your cooperation and saying what I intend, my self-understanding will be far simpler and more stable than any instrumental understanding I can achieve in these circumstances. Unlike an instrumental understanding of my behavior in this case, an expressive understanding will not undermine itself, suspicions to the contrary notwithstanding. The thought that instrumental calculations are revived at the prospect that I might be interpreted as thinking expressively and hence as sincere – *that* thought occurred to me just now, not in my imagined capacity as an agent thinking expressively about his behavior in a prisoners' dilemma, but rather in my capacity as a philosopher accommodating his reader's bias in favor of instrumental thinking. As an agent thinking expressively about his behavior in a prisoners' dilemma, I would find a perfectly stable self-understanding in the conception of myself as intending to cooperate and expressing my intention. Expressive thinking would not itself lead back to instrumental calculations, and if it did, those calculations would be unstable, as we have seen.

In short, I face a choice antecedent to the choice between sincerity and insincerity – namely, the choice between thinking instrumentally and thinking expressively about that subsequent choice.[14] Thinking

[14] Here my argument is similar in form to David Gauthier's argument about the choice between straightforward and constrained maximization. But there is a crucial difference between us. According to Gauthier, an agent chooses between straightforward and constrained maximization as the fundamental principle of his practical reasoning, and so he

instrumentally leads to an endlessly vacillating calculation, whereas thinking expressively leads to a clear and consistent self-understanding. Whether honesty or dishonesty is the best policy, in the sense of yielding the best consequences, is a vexed question whose answer defies deduction. But honesty is certainly the clearest, most perspicuous policy, the policy that affords me the clearest sense of what I am about. I think that many of us adopt the policy of honesty on precisely these grounds.

I do not claim to have shown that the rational pressure in favor of sincerity always prevails. In particular, there are extreme losses that it makes sense to take a shot at avoiding, and extreme gains that it makes sense to take a shot at obtaining, no matter how wild or how blind a shot. But there are many gains and losses that it makes more sense to ignore, given the more intelligible alternative of speaking our minds. And what's more intelligible is, on my view of practical reason, the more rational course to take.

Thus far I have addressed only the problem of credibility – of how one agent might give another valid grounds to believe that he has formed an intention to cooperate. I haven't yet addressed the problem of commitment. How can an agent form a cooperative intention that will take effect in his future behavior, given the incentives for his future self to change his mind?

In discussing Anscombe's theory of intention, I confined myself to immediate intentions to act. This discussion is not immediately relevant to the intentions required by cooperative agreements of the sort that would offer an escape from the prisoners' dilemma. The latter must be long-range intentions, to do something in the future, when the relevant conditions have been fulfilled and the opportunity arises. Such long-range intentions do not appear to offer any shortcut to the cognitive aim of knowing what I am doing here and now.

can choose only on the basis of whichever one of these principles he last chose. There is no prior, unchosen principle with which to reason about his choice. In my view, however, the agent chooses between instrumental and expressive thinking, not as fundamental modes of practical reasoning, but as different versions of the one mode of thinking that constitutes practical reasoning, antecedently to his choice – namely, making sense of what he does, by doing what makes sense. What shows that this practical pursuit of self-knowledge constitutes practical reasoning is, not that a rational agent would choose it, but that it helps us to explain many of the phenomena of rational agency, including the nature of intention, an agent's non-observational self-knowledge, and so on.

What long-range intentions do, of course, is enable me to coordinate my behavior at different times – to take present steps in preparation for future steps that I am going to take, to postpone steps until later so that I needn't take them now, to think through at leisure a sequence of steps that I will have to execute when there's no time to think.[15] Yet this instrumental function of long-range intentions rests, at bottom, on the cognitive function of letting me know what I am going to do in the future. In order to take a trip next month, I must buy a plane ticket in advance, but I can see no reason for buying the ticket until I know that I would indeed use it to take the trip. The intention to take the trip gives me access to a reason for buying the ticket, by ruling out the possibility of its going to waste. Similarly, intending to buy the ticket this evening can cancel my reason for taking out my cell phone and buying it right now – a reason that conflicts with my reasons for finishing this essay. My reason for buying the ticket now is cancelled by the knowledge that I needn't do so, because I am going to buy it later.

Furthermore, the ability to know what I am going to do in the future enables me to know what I am doing now in terms that are even more comprehensive or integrative than before. I am not just writing an essay and postponing the purchase of a ticket; I am postponing the purchase of a ticket so as to finish the essay that I am due to present at the conference to which I will buy a ticket later this evening. My present action can therefore be understood as one step in a temporally coherent course of action, but only because I can expect to take the future steps with which it will cohere.

The knowledge embodied in my long-range plans bears several points of resemblance to that embodied in my immediate intentions. For one thing, it depends for its possibility on my having an objective conception of myself as one of the world's creatures, toward whom I occupy an epistemic position somewhat similar to yours. The question what I am going to do in the future simply wouldn't arise for me if I couldn't conceive of a future person who would be me.

What's more, my epistemic position with respect to this future person affords me a shortcut to knowledge about him. In order to take my present intentions for the future as predictive of future action, I must

[15] This sentence summarizes many of the points made by Michael Bratman in *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987). I discuss Bratman's view at greater length in "What Good Is a Will?" (MS).

have grounds for expecting them to be fulfilled, but I am fortunately in a position to give myself those grounds, by fulfilling my past intentions for the present and thereby demonstrating my tendency to fulfill long-range intentions. My intellectual drives therefore favor fulfilling my past intentions and can be expected later to favor fulfilling my present ones. When those drives are directed toward my objectively conceived self, they motivate me, not only to be intelligible to myself, but also to give myself evidence of my own reliability.[16]

What generates this rational pressure toward fulfilling commitments is that, although the present dilemma is the first and last one that I will ever share with you, it is not the only one that I share with myself. I have no incentive to convince you that I tend to reciprocate cooperation, because I will have no opportunity to realize the benefit of that conviction on your part; but I do have an incentive to convince myself that I tend to carry out my long-range intentions, because my ability to settle what I will do in the future depends on my grounds for that conviction.[17]

[16] I am not imagining here that constancy, as I call it, is a distinct disposition – a disposition specifically to carry out intentions – on which I must rely when forming intentions. If it were, then carrying out intentions for the purpose of giving myself evidence of that disposition would be self-deceptive, since it would manifest my desire for that evidence rather than the distinct disposition of constancy. As I imagine it, however, constancy can consist in any psychological state or mechanism that makes me reliable in fulfilling intentions. My constancy can even consist in my desire for the ability to tell what I am going to do in the future, if that desire motivates me to fulfill my intentions. What I want, after all, is some grounds or other on which I can regard my future course of action as determined, even if those grounds consist in the fact that this very desire will determine me to do what I have regarded in that way. And there is nothing self-deceptive about being motivated by this desire to carry out intentions for the sake of giving myself evidence that I am so motivated. The thought of fulfilling an intention in order to maintain my grounds for relying on my own intentions is not undermined by the realization that I would be fulfilling the intention for that purpose, since that purpose is one I can rely on myself to have. I discuss this mechanism in greater detail in Chapter 8 of *Practical Reflection* (Princeton: Princeton University Press, 1989).

[17] Another way of making this point is to note that Gregory Kavka's famous toxin puzzle can be solved by iteration. (See Kavka, "The Toxin Puzzle," *Analysis* 43 (1983), 33–36.) The puzzle is this: A mind-reader offers to give you a million dollars if you form the intention to drink a toxin that will make you painfully ill, without causing any lasting injury. The mind-reader will pay you the money as soon as he detects your intention, leaving you with an intention that you have every reason not to fulfill. Can you form the intention, and if you do form it, should you fulfill it?

In my view, an intention to drink the toxin would entail a cognitive commitment to the truth of the proposition that you are going to drink it – a commitment of the sort that would constitute knowledge if it were true and appropriately justified. Because of being an intention rather than a mere prediction, however, this commitment would have to be justified in part by its own power to bring about the facts that would make it true. Thus, you must form the intention to drink the toxin by committing yourself to

Finally, my access to this epistemic shortcut is known to you and can give you grounds on which to reason about my behavior. Just as you realize that I share your interest in finding me intelligible, so you realize that I share your interest in finding me reliable, because I conceive of myself as if from your perspective, as a creature from whom either unintelligibility or unreliability would be problematic. Knowing of my need to understand my behavior, you are entitled to interpret my offer of cooperation as the genuine expression of an intention, which would be intelligible, rather than a strategic gambit, which would not. Knowing of my need to project my future behavior, you are entitled to expect me to carry out my cooperative intention, in order to preserve my grounds for such projections.[18]

the truth of the proposition that you will drink it; and you cannot commit yourself to the truth of the proposition unless you can expect thereby to cause the proposition to come true.

Now suppose that Kavka's mind-reader offers to play the toxin game with you many times, and suppose that you succeed in forming the crucial intention on the first play. In that case, you will realize that fulfilling your intention, by drinking the toxin, is essential to maintaining your ability to form similar intentions on future plays. For if on the second play you knew that, in the only relevant prior instance, you formed the intention but then failed to fulfill it, then you would not be in a position to commit yourself to the truth of the relevant proposition in this present instance, because you would not be in a position to expect thereby to cause the proposition to come true; and so you would be unable to form the second intention and claim the second prize. So when you have formed the intention and collected the prize on the first play, you will see that fulfilling your intention is essential to preserving your ability to form the intention and claim the prize on subsequent plays. Rationality will therefore favor drinking the toxin.

The only reason why the toxin puzzle is puzzling to begin with is that the situation is described so as to seem overwhelmingly unlikely to recur. If the situation were described in terms that highlighted its similarity to everyday situations that call for resoluteness, it wouldn't seem so puzzling, since the importance of retaining grounds for planning would be clearer.

[18] According to this conception of the reasons generated by an intention, their strength can vary with the circumstances. In circumstances of some kinds, the agent doesn't especially want or need the ability to settle the question what he is going to do on some future occasion. If he does settle the question in some particular instance of such circumstances, by forming an intention, he will feel especially free to unsettle it again, by reconsidering or changing his mind, since a record of inconstancy is of little consequence in circumstances of that kind. In circumstances of other kinds, however, the agent really does need to know and be able to say what he is going to do on some future occasion, and a record of inconstancy in those circumstances would be seriously problematic. An agent has greater motivation for vindicating his own self-trust when circumstances are of the kind in which the ability to tell what he is going to do is especially important. For that very reason, however, he has better grounds for self-trust in circumstances of this latter kind, knowing that he will have motives for rising to the occasion. And circumstances of this kind surely include the opportunity to escape a prisoners' dilemma through a cooperative agreement.

Conclusion

A person who gives himself no grounds for credence in his long-range intentions, or who gets tangled up in instrumental reasoning about the truth, sacrifices a considerable degree of self-knowledge. The objectively conceived personality to which this person has pinned his subjective point-of-view is less intelligible and less predictable than it otherwise might be.[19] In this respect, the axis on which he is centered, or the anchor by which he is grounded, is less sure. A person who says what he thinks and does what he says has a better grasp on the person who he is. He can therefore be described as better centered or better grounded.

I have now arrived at the explanation that I set out to find, of the relation between being centered and being trustworthy. The relation is that the trustworthy person has a surer sense of self than the person who strategizes with the truth or defaults on his commitments.

As I mentioned earlier, I have purposely developed this explanation in abstraction from the social, emotional, and moral considerations that bear on prisoners' dilemmas in real life. My reason for adopting this idealization has been to isolate rational pressures toward trustworthiness within the individual perspective of a rational agent considered merely as such. I do not believe that these pressures are sufficient in themselves to resolve actual dilemmas. Rather, I believe that they subtly favor the gradual accretion of social, emotional, and moral resources that jointly provide a resolution. Explaining this process is more than I can do in this essay, but let me offer an idea of how that explanation would go.

The process of practical reasoning, as I conceive it, extends beyond the immediate step of doing what makes sense here and now. I have already mentioned one further step, in which one fulfills intentions from the past in order to preserve the credibility of one's intentions for the future. In my view, there are many other ways by which one can cultivate intelligibility in oneself. For example, one can try to resolve conflicts among one's ends, so as to avoid situations in which one would have trouble explaining the pursuit of either end given one's commitment to the other. One can also adopt policies of behavior that generalize about how one deals with situations of repeatable kinds. A particularly fruitful

[19] Note that the predictability at issue here is not of the boring sort that characterizes a person set in his ways. The predictability at issue is that of a person who is in a position to know what he will do in the future precisely because he is in a position to make it up, by making up his mind.

kind of policy – fruitful, that is, for self-understanding – is a norm of the sort described by Allan Gibbard.[20] Accepting such a norm involves adopting a disposition to favor or oppose the relevant kind of behavior, by adopting or eschewing it oneself and approving or disapproving of it in others – a broad pattern of conduct that can be understood in terms of a single attitude. Finally, one's conception of oneself will gain in generality and explanatory power insofar as it can be subsumed under one's conception of people in general – who are, of course, similarly striving to understand themselves under self-conceptions subsumable insofar as possible under a conception of people in general, including oneself. People are therefore jointly encouraged to converge on a conception of what "we" are like, or how "we" live, so that they can understand themselves individually, to some extent, by conceiving of themselves as one of "us."

In pursuing these long-range strategies of practical reasoning, one is influenced by the cognitive attractions of saying what one thinks and doing what one says. For example, cultivating ends and norms compatible with truth-telling, and weeding out ends and norms incompatible with it, will enable one to avail oneself of expressive self-understanding without any confusing motivational conflict. That's how the fairly subtle pressures that I have identified in the perspective of the bare rational agent can lead to the gradual accretion of additional resources for coping with prisoners' dilemmas in real life. I believe that a fuller exploration of this process would yield a detailed explanation of why it is rational to be a *Mensch.*

Appendix A: Perry's Theory of Self-Knowledge

My opening remarks about the differences between humans and cats were based in part on Thomas Nagel's theory of the "objective self" and

[20] *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, MA: Harvard University Press, 1990). Note that my use of Gibbard's idea differs from his in one crucial respect. According to Gibbard, different consistent sets of norms can be comparatively assessed only in light of some higher-order norm, which itself is evaluable only in light of some yet higher-order norm, and so on. In my view, however, practical reasoning has a substantive criterion of success – self-understanding – in light of which alternative sets of norms can be assessed. The foundation for this criterion of rationality lies, not in our adoption of some norm, but rather in the nature of autonomous action, as revealed by moral psychology. On this last point, see my paper "Deciding How to Decide," in *Ethics and Practical Reason,* ed. Garrett Cullity and Berys Gaut (Oxford: Clarendon Press, 1997), 29–52.

John Perry's theory of self-knowledge.[21] In this Appendix, I summarize Perry's theory and expand upon my application of it.

Perry begins with a form of self-knowledge that he calls "agent-relative," which characterizes the world in terms that are implicitly relative to the subject. For example, my belief that there's an accident blocking the road up ahead is true if and only if the accident is situated ahead of David Velleman, and in that sense the belief is about myself. Yet this aspect of its truth-condition is not explicit, either in my verbal expression of the belief or in the mental representation that is a constituent of the belief itself. That is, I do not refer to myself as the person ahead of whom the accident is situated, nor do I exercise an idea of myself in mentally representing the accident as up ahead. In this respect, my verbal expression of the belief and the belief itself are elliptical.[22]

What implicitly fills the ellipsis is indicated by the point-of-origin in my perspective. A visual image, for example, is organized along sight-lines that converge on a point presumably occupied by the unseen subject of vision; and things are represented in the image as "up ahead" by being implicitly represented as ahead of that point and its presumed occupant. Similarly, an utterance presumably issues from the mouth of a speaker, and what is represented as "up ahead" in the utterance is implicitly represented as ahead of the presumed speaker.[23] When I believe that there is an accident up ahead, the content of my belief must likewise be framed from a perspective, and the origin of that perspective must be the point to which I implicitly believe the accident to bear the relation "up ahead."

[21] See John Perry, "Self-Notions," *Logos*, 1990: 17–31; and "Myself and 'I'," in *Philosophie in Synthetischer Absicht* (A Festschrift for Dieter Heinrich), ed. Marcelo Stamm (Stuttgart: Klett-Cotta, 1998), pp. 83–103. See also "The Problem of the Essential Indexical," *Noûs* 13 (1979): 3–21.

[22] On egocentric thought that is elliptically first-personal, see also D. H. Mellor, "I and Now," *Proceedings of the Aristotelian Society* 89: 79–94 (1989), reprinted in *Matters of Metaphysics* (Cambridge: Cambridge University Press, 1991); and the discussion of Mellor in José Luis Bermúdez, *The Paradox of Self-Consciousness* (Cambridge, MA: MIT Press, 1998), Chapter 2, section 2.1. Bermúdez believes that the notion of elliptically first-personal thought must be defended against "the classical theory of content," according to which a subject cannot have thoughts without having concepts sufficient to compose complete propositions to serve as their contents. While I agree with Bermúdez in rejecting this theory of content, I do not think that we need independent grounds for rejecting it – grounds independent, that is, of the obvious counterexample consisting in elliptically first-personal thought. (Nor do I agree with Bermúdez's view that, once we reject the classical theory of content, we must fashion a positive theory of nonconceptual content in order to account for first-personal thought.)

[23] Of course, if a different point-of-view is more salient in the context than the speaker's, then the words "up ahead" are interpreted in relation to that point-of-view. My point here is merely that the speaker's perspective is the default.

Perry replaces the notion of a perspective with that of an "epistemic/pragmatic relation" – that is, a relation that structures ways of detecting things and ways of dealing with them, which Perry describes, in turn, as epistemic and pragmatic methods. To represent a traffic accident as "up ahead" is to represent it in a manner that's structured like the output of epistemic methods for detecting what's up ahead, and like the input to pragmatic methods for dealing with what's up ahead. The former methods include looking into the middle distance in the direction I am traveling and switching on my headlights; the latter include honking my horn to alert people blocking the road; and these methods can be combined, as when I take a second look to see whether a second honk is needed. The "up ahead" relation determines how my representations of the world are structured when obtained by the former, epistemic method and how they must be structured in order to guide the latter, pragmatic method.

Some epistemic/pragmatic relations are reflexive: they necessarily structure information that is received from, and relevant to dealing with, the agent's own body or mind. My proprioceptions and tactual sensations arrive as if from particular locations, which I recognize by their orientation in tactual or proprioceptive space; and I direct muscular control at locations similarly oriented in kinaesthetic space. These orientations can be expressed by phrases such as "my right hand" or "my left foot," but the fact that they pick out a hand or a foot is not evident in the orientations themselves. That is, sensations felt "in my left foot" are not felt as originating in a particular anatomical structure; they are simply felt as "there," in tactual space, under an epistemic/pragmatic relation that is ultimately ineffable. Similarly, I wiggle my left toes by wiggling "there," a location conceived under the same ineffable relation.

What's accessible under reflexive epistemic/pragmatic relations is often accessible under relations that are not reflexive. The source of sensations felt "there" (in my right hand), which also moves when I move "there" (with my right hand), is an object that is seen as "to my right," under a relation that isn't reflexive, because it can be occupied by many things that aren't part of my body, if they come to occupy that region of my visual field. My dual relation to such objects allows for a rudimentary kind of self-knowledge that can be formulated entirely within agent-relative thought. If I see a tangle of arms to my right, then making a movement with my right hand may reveal an important fact – namely, which of the visually perceived arms is mine.

This sort of self-knowledge is even available to my cat Snowflake – for example, when she recognizes that the tail she is chasing is her own.

When Snowflake sees her tail, she sees it "to the rear," a relation that also governs epistemic methods of hearing as well as pragmatic methods of fleeing and chasing. "To the rear" is a very different relation from "my rear end," which governs epistemic methods of tactual sensation as well as pragmatic methods of flicking and licking. When Snowflake sees her tail merely "to the rear," she may end up chasing it; she finally stops chasing it when she connects what she is seeing rearward with what she is feeling and doing rear-endward. She can then be said to have recognized that the tail she is chasing is her own.

But in what sense does Snowflake recognize the tail as her own? All she does is forge a mental association between an object seen "back there" in visual space and the locus of sensations felt "there" in tactual space, or of movements made "there" in kinaesthetic space, so that she can now think of causing the thing "back there" to wave by waving "there," or of causing a sharp sensation "there" by nipping the thing "back there." To be sure, her sensory-motor relation to "there" is reflexive, because it picks out a location within her own body; and its being reflexive in this sense is our grounds for crediting her with the discovery that the tail is her own. But Snowflake remains unaware that "there" is a location within her body, because she does not conceive of herself as an embodied subject. She is unaware that sensations felt "there" and movements executed "there" are the perceptions and actions by which the conscious life of a particular creature extends to a part of its body; and so in recognizing the object seen to her rear as the locus of those sensations and movements, she does not conceive of it as belonging to herself in the way that a tail belongs to its owner. She has no conception of a self to whom the object seen rearward might belong – of a creature who she is and who feels sensations and executes movements with that object.

The theoretical tools presented thus far will therefore have to be supplemented if they are to account for the self-knowledge that separates man from cat. Perry supplements them with the notion of "self-attached knowledge."

Compare two different ways in which I can recognize my reflection in a store window. First, I can associate the movements made by a figure reflected in the window with the movements that I am aware of making; second, I can recognize the same reflected figure, by his looks, as me. The first recognition is of the sort that the cat attains when she recognizes her own tail: it involves no more than making a connection between non-reflexive and reflexive perceptions, all of which are still agent-relative.

But the second recognition requires me to have standing non-reflexive knowledge about myself: I have to recognize a particular appearance as mine.

Now, my knowledge that a particular appearance is mine cannot depend on its being structured in a way that's distinctive of reflexive methods, since the relevant appearance is structured by a perspective other than my own. It's the appearance I have when seen as another person. How do I remember that this appearance as of another person is mine?

Perry's answer is that I establish a more lasting association between reflexive and non-reflexive information about myself. I frame a standing idea of a human body, and I use it to store, as attributable to that body, information about what I reflexively feel "in my body" and do "with my body," and as a source of information to guide such reflexive methods. Because it is thus associated with my reflexive methods, the idea comes to represent the particular human body that is mine; yet because it is framed from no particular perspective, it can serve as my repository for non-egocentrically structured information about my body, such as information about how it looks from perspectives other than my own. This information will be marked as pertaining to my own body, not by virtue of the structure of its representation, but by virtue of being stored in an idea that is permanently associated with my own-body–oriented methods.[24] That's how non-egocentric knowledge about the person I happen to be can become, as Perry puts it, self-attached.

Perry describes this idea as my conception of "the person identical," explaining that the one to whom the person is therein represented as identical is the unrepresented subject. "The person identical" is elliptical, because it doesn't specify to whom the person is identical; but in that respect, Perry argues, it is on a par with "up ahead," which doesn't specify the anchor of that relation, either.[25]

I would add one final note to Perry's analysis, which seems to end prematurely. To conceive of a person as the one whose left foot is the locus of what I feel "there" (in my left foot) or do "there" (with my left

[24] In order to become my idea of myself as a whole person, this idea would have to be associated with my reflexive methods of introspection and thought-control, so that it incorporated information about my mental states as well.

[25] As Perry points out, specifying the anchor of the relation "the person identical" would lead to a vicious regress. Indeed, avoiding this regress is the purpose of positing elliptically first-personal thoughts, in the first place. See also Mellor, "I and Now," and Bermúdez, *The Paradox of Self-Consciousness*, loc. cit.

foot) is not yet to conceive of that person as the "I" feeling and doing those things. The latter conception would have to represent the person not only as the target but also as the subject of my reflexive epistemic/pragmatic methods, and so it would have to be associated with my reflexive methods more closely, as follows. What I feel "in my body" or do "with my body" must be represented in this idea of a person, not just as being felt or done in that person's body, but also as being felt or done by that person. Hence this idea of a person must represent him as using reflexive methods to detect and cause the events that I detect and cause "there," in and with my body. So conceived, that person will fully occupy the role of the person identical, and my conception of him will be a conception of who I am – or, as it is often called, a sense of identify or self.

Appendix B: Reasons for Acting

My conception of reasons for doing something is that they are considerations in light of which one's doing that thing would make sense, because they would help to explain one's doing it. This conception of reasons raises various objections, to which I have offered replies in my first book, *Practical Reflection*, and in a symposium on my second book, *The Possibility of Practical Reason*.[26] In this Appendix, I summarize a few of those objections and replies.

The most obvious objection is that reasons for acting are not about oneself and one's attitudes, as I claim, but rather about those aspects of the world at which one's attitudes are directed. This objection depends, I believe, on a confusion between the logic of practical reasoning and the explicit content of practical thought.

If you look up from reading *Felix Holt* and say to yourself, "What a genius she was!" your thought is explicitly about the author George Eliot; but in articulating this thought, you express an attitude that lends intelligibility to various further thoughts and actions on your part. Suppose that your next thought is "I wonder what else she wrote" (or perhaps just "What else did she write?"). The rational connection between your thoughts is that admiration of the sort expressed in the first naturally leads to curiosity about its object, as reported (or expressed) in the second. This connection cannot be discerned in the explicit content of your thoughts. There is no rule of inference leading from the premise that George Eliot was a genius to the conclusion that you wonder what she wrote in addition

[26] *Philosophical Studies* 121 (2004): 225–38, 277–98.

to *Felix Holt.* Unless the first of these thoughts is understood as expressing an attitude held by the thinker of the second, they amount to a *non sequitur.*

The only way to make the logic of these thoughts explicit would be with further, reflective information – "I admire the author of *Felix Holt* as a genius, and so I am moved to wonder what else she wrote" – which describes a psychologically intelligible transition of thought. Yet to articulate this reflective information to yourself would be to shift the focus of attention, from the author whom you admire to your own attitude of admiration. And this shift would make your admiration less rather than more evident, because admiring someone entails attending to her rather than yourself. "I admire the author of *Felix Holt*" would be a less admiring thought, a thought less expressive of your attitude, than "She was a genius." Articulating your awareness of admiring Eliot would therefore leave you less vividly aware of admiring her than articulating thoughts expressive of that admiration, which would be thoughts about Eliot.

Thus, explicit reflection is often self-defeating. Reflective reasoning is best left implicit, in the background, so that the attitudes that are its objects can be revealed more clearly in explicit thoughts about other things. Hence the fact that your thoughts prior to acting are not explicitly about yourself is no evidence that their logic is not reflective. Thoughts that are explicitly about other things may yet be structured by what they reveal about yourself – as in "What a genius she was! I wonder what else she wrote."

Note that this response to the present objection points to a flaw in the traditional philosophical method of studying practical reason. The traditional method is to construct an argument-schema that will both represent the explicit content of, and illustrate the rational connections among, the thoughts leading up to an action performed for reasons. Aristotle's practical syllogism was the first attempt to construct such an argument-schema, and many other attempts have followed. In my view, however, the rational connections in an agent's deliberations are connections of reflective intelligibility, and such connections tend to hold, not between the contents of the agent's explicit thoughts, but rather between the self-attributions that remain in the background, implicitly registering the attitudes that his explicit thoughts express. Because these unarticulated self-attributions provide the logical structure of the agent's thinking, they contain the agent's reasons for acting, in my view; but because they remain unarticulated, they cannot be represented by the same argument-schema that represents the agent's explicit thinking.

In sum, an agent's reasons for acting are not the things that he says to himself before acting. That he doesn't say anything about himself to himself before acting doesn't prove that his reasons for acting are not considerations conducive to self-understanding.

Another objection to my view of reasons for acting is that an agent may understand his behavior in terms of unfortunate traits that do not provide reasons for their behavioral manifestations. Someone who knows himself to be lazy, for example, may find his avoidance of work intelligible in that light without thereby finding it supported by reasons. My answer to this objection is that the conception of himself as lazy, rather than as easygoing or laid back, expresses disapproval, which would have to be included in a complete conception of himself. And manifesting laziness while condemning it as such is not altogether intelligible, after all.

A deeper objection is that although my account explains the influence of reasons, it fails to explain their normative force.[27] My answer to this objection has two parts. First, the intellectual drive that reasons for acting engage, in exerting their influence, carries a kind of authority by virtue of being inextricably identified with the agent himself. The agent cannot stand back from his drive toward self-understanding and regard it as an alien influence on him, because regarding it as an influence at all is an exercise of self-understanding, animated by the self-same drive, which consequently has not been banished to the realm of the alien, after all.[28]

The second part of my reply to the present objection is that the normative force of reasons for acting may be supplied to some extent by a norm in favor of doing what makes sense – a norm that we adopt in the course of pursuing self-knowledge, precisely because it helps us to make sense of that very pursuit. Practical reasoning, as I conceive it, favors the adoption of norms that ratify and regularize aspects of our behavior. When norms are accepted consciously, they provide generalizations that guide our behavior by offering us the means to understand the behavior so guided. In adopting the posture of being "for" some things and "against" others, we thereby adopt a comprehensive description for some

[27] This objection has been pressed independently by Nishi Shah, Kieran Setiya, Nadeem Hussain, and Matthew Silverstein. See Hussain's contribution to the *Philosophical Studies* symposium on *The Possibility of Practical Reason* and Shah's paper "How Truth Governs Belief," *The Philosophical Review* 112(2003): 447–82.

[28] This point is developed further in "What Happens When Someone Acts?" in *The Possibility of Practical Reason* (Oxford: Oxford University Press, 2000), 123–43; and in "Identification and Identity" (Chapter 14 in the present volume).

region of our conduct, which subsequently tends to follow suit, so as to be comprehensible under that description. We adopt the norm of doing what makes sense in order to regiment and make sense of a process by which our actions are already regulated – in this case, the very process of making sense of what we do by doing what makes sense. Hence the natural process of attaining practical knowledge affirms itself, by leading to the adoption of a norm that ratifies and regularizes it as the process of practical reasoning.[29]

[29] This paragraph borrows significantly from unpublished work by Nishi Shah.