The Birth of Ethics

PHILIP PETTIT

The Tanner Lectures in Human Values

Delivered at

University of California, Berkeley
April 7–9, 2015

PHILIP PETTIT is L. S. Rockefeller University Professor of Politics and Human Values at Princeton University, where he has taught political theory and philosophy since 2002. Since 2012–13 has held a joint position as Distinguished University Professor of Philosophy at the Australian National University, Canberra. Born and raised in Ireland, he was a lecturer in University College, Dublin, a research fellow at Trinity Hall, Cambridge, and professor of philosophy at the University of Bradford, before moving in 1983 to the Research School of Social Sciences, Australian National University. There he held a professional position jointly in Social and Political Theory and Philosophy until 2002. Prof. Pettit was elected fellow of the American Academy of Arts and Sciences in 2009, honorary member of the Royal Irish Academy in 2010, and corresponding fellow of the British Academy in 2013. He has also been awarded numerous honorary professorships and degrees.

Prof. Pettit works in moral and political theory and on background issues in the philosophy of mind and metaphysics. His single-authored books include *The Common Mind* (Oxford University Press, 1996), *Republicanism* (Oxford University Press, 1997), *A Theory of Freedom* (Oxford University Press, 2001), *Rules, Reasons and Norms* (Oxford University Press, 2002), *Penser en Société* (Presses Universitaires de France, 2004), *Examen a Zapatero* (Temas de Hoy, 2008), *Made with Words: Hobbes on Mind, Society and Politics* (Princeton University Press, 2008), *On the People's Terms: A Republican Theory and Model of Democracy* (Cambridge University Press, 2012), *Just Freedom: A Moral Compass for a Complex World* (Norton, 2014), and *The Robust Demands of the Good: Ethics with Attachment, Virtue and Respect* (Oxford University Press, 2015).

INTRODUCTION

FROM REDUCTION TO GENEALOGY

According to the ethical point of view, as commonly understood, there are two striking aspects to the natural and social world we human beings inhabit. First, the options we confront in decision making vary in their overall desirability, however desirability is conceptualized. And second, we are often fit to be held responsible in making a choice for choosing or failing to choose the most desirable option. Specifically, we are fit to be held responsible for the option we choose in the presence of a capacity—an unimpaired, unimpeded capacity—to register and act on considerations of desirability. Depending on what we do, we are appropriate targets for the praise and blame of others and can appropriately feel pride or guilt in how we perform.

The concepts of the desirable and the responsible are essentially prescriptive or normative. To hold that one option in a given choice is more desirable than alternatives is to prescribe its performance, on the assumption that other things are equal. To hold that an agent is fit to be held responsible for the choice is to maintain that other things are equal and that it is appropriate, therefore, to prescribe the most desirable option in advance or, depending on what was actually chosen, to deem the choice praiseworthy or blameworthy in retrospect: to commend or condemn the action or, if you are the agent, to feel pride or guilt about how you behaved.

Because of being inherently prescriptive, ethics raises a problem for those of us who think that the world we live in is an austere place that conforms to the image projected in natural science and mathematics. For prescriptive properties like desirability and responsibility—that is, fitness to be held responsible—do not look to be of a kind with mathematical properties.[1] And neither do they seem to be at home in the naturalistic world of science. They are unlikely to pull weight in any of the laws that science seeks to identify, and they do not plausibly materialize in virtue of effects that those laws explain.

Naturalistic philosophers have sometimes responded to the problem raised by debunking the idea that, by naturalistic lights, desirability and responsibility are bona fide properties. They have opted for representing ethical talk as fundamentally emotive or expressive, for example, rather than taking it to be descriptive of any features of items in the world.[2] Or they have held that in speaking ethically we treat desirability and

responsibility as if they were real properties when actually they are not: consciously or otherwise, so this story goes, we operate with a fiction.[3]

The idea in these lectures is to resist downgrading ethical discourse in any such manner and, without forsaking naturalism, to try to vindicate the assumption that there really are properties like desirability and responsibility in the world and that they have an impact on our actions. These properties are not on a par with natural properties like mass and charge and spin, of course, which characterize the fundamental building blocks of our universe. The assumption is that just as the pixels on a television screen support patterns or properties at a higher level—the patterns we register in following any TV program—so the patterns we discern in distinguishing desirable actions and responsible agents are supported at a lower level by the fundamental elements—ultimately, perhaps, wavicles or strings—out of which the world is built.

The standard way to vindicate this sort of realism about higher-level, ethical properties, at least within a scientifically based view of the universe, is to try to provide a naturalistic reduction of ethical talk. Such a reduction would argue, first, that there is nothing more to the realization of an ethical property than the realization of a suitably supportive natural configuration—there may be an endless variety of these—as there is nothing more to the realization of a pattern on a TV screen than the realization of a suitably supportive configuration of pixels. And it would try to show, second, that whatever facts we register about the ethical property, or indeed the property presented on the TV screen, we might in principle have registered by noting how things stand at the lower, supportive level. It would maintain, roughly, that there is a sense—there are many candidates for what this sense is—in which the higher-level language can be reduced to lower-level terms.[4]

These lectures explore a distinct way of vindicating a naturalistic realism about ethical properties like desirability and responsibility. The idea is not to argue for a particular way of reducing ethical talk to naturalistic talk but to provide a naturalistic genealogy of how ethical talk could have arisen, in particular a genealogy under which ethical judgments play a role in registering bona fide aspects of the world and in shaping our responses to that world. The aim is to vindicate ethics, taken literally or realistically, in naturalistic terms. And the plan is to achieve that aim by explaining how we, the products of a natural and cultural evolution, could have come to develop notions of desirability to refer to aspects of

the options we face, to shape our choices between those options, and to determine our fitness to be held responsible for what we do.[5]

To vindicate a naturalistic realism about ethics, establishing that there are bona fide properties like desirability and responsibility in the world, is not to deny that those properties may be inherently anthropocentric. The exercise pursued here takes those properties to consist in patterns that become visible, and only become visible, from within a perspective that presupposes access to distinctively human practices. The anthropocentric character of the properties identified does not argue against construing them, however, in a naturalistic, realist fashion. The colors that we perceive on the surfaces of objects are detectable, and perhaps only detectable, from within the sort of visual processing systems we and our biological ilk bring to the world. But their anthropocentric character does not, or should not, lead us to reject a naturalistic realism about such colors.

A CONJECTURAL HISTORY OF ETHICS

For purposes of these lectures, vindicating ethics can be taken henceforth to mean vindicating a naturalistic realism, however anthropocentric, about desirability and responsibility. A genealogical vindication would start with a possible, naturalistically intelligible form of human society where people do not have access to ethical concepts, and then show how naturalistically intelligible adjustments would lead them to develop and deploy such concepts in charting their world. It would amount to a conjectural history of ethics, as it might have been described in the eighteenth century. This does not aim at a conjectural narrative about how ethics really emerged—the goal is not a just-so story—but at an explanation of how ethics would really have emerged under certain conjectural conditions.[6] The aim is to establish the naturalistic emergability of ethics.[7]

I start in this exercise with a naturalistically plausible, if historically unlikely social state where members lack an ethics or morality; they do not have access to the network of practices and concepts associated with desirability and responsibility. And I then try to show that people in that state would plausibly have had motives and opportunities sufficient to push them onto a trajectory of development culminating in the ethical. They would have been led, as by an invisible, nudging hand—and not, for example, as a result of foresight and planning—to invoke standards of desirability, and to hold one another responsible for living up to those standards.

In the social state imagined at the origin of this development, natural language has already emerged, presumably on the basis of a naturalistically intelligible process of natural and social evolution. People use that language, however, only for purposes of giving one another reports on how things are in their environment, according to their own beliefs: whether the blackberries have ripened on the hill, what the weather is like farther north, how the prospects are looking for a big-game hunt. In particular, they make no ethical pronouncements bearing on issues of what it is desirable to do or on who is fit to be held responsible for something done.

Anticipating later discussion, the claim is that once people can use words to communicate their representations of the world in this manner, they are more or less bound to develop further speech acts of avowal and pledging, co-avowal and co-pledging, and to put themselves thereby in a world where ethical practices and concepts can gain traction. Or at least they are bound to do this, on the assumption that they display a variety of characteristically human features. However culturally malleable, for example, they are disposed by nature to exercise joint attention, consciously focusing on data they take to be available to all, albeit from different perspectives[8] However altruistic in other ways, they are deeply invested in promoting their own welfare and that of their kin. And however individually resourceful, they need to establish and maintain relationships and networks of mutual reliance in order to promote that welfare: they need to be able to rely on others and to get others to rely on them.[9]

It is unlikely that there ever was a time or place in the trajectory of human development when our ancestors used language solely for making reports on their shared world. And it is even more unlikely that a society at any place or time would have existed in isolation from other societies, as simplicity requires us to assume here. For this reason, I use the name of Erewhon to refer to our starting society. This name, borrowed from a nineteenth-century novel, is an anagram of "nowhere" and may serve to remind us of the unhistorical nature of the community with which the narrative begins. We may think of Erewhon as a possible scenario rather than as an earlier stage in human history and treat the narrative as an exploration of how ethics would be liable to emerge in that possible world.

Our narrative about Erewhon is also unfaithful to history in assuming an equality of power that discounts rigid hierarchies of gender or class; it projects a picture of communicative exchanges in which power and domination play no role. Our species has been on Earth for at least a

hundred thousand years and we know that since the agricultural revolution that occurred about seven or eight thousand years ago, inequality of power has been the rule, not the exception. In supposing a relatively egalitarian Erewhon, then, the narrative does not reflect recent human history.

This particular inaccuracy need not be very troubling, however. There is some ground for thinking that preagricultural societies were much more egalitarian than agricultural, so that it is not clear how far we are departing from history on this front.[10] And in any case a departure from history on that front would not be a problem for the enterprise undertaken here. It would scarcely be a strike against the naturalistic intelligibility of ethics that suitable practices and concepts could only have arisen naturalistically in an egalitarian community; that might teach a lesson about the nature of ethics but it would hardly put its naturalistic credentials in doubt.

But not only is the starting point in our narrative unhistorical; more importantly, the process invoked in the account of how ethics could emerge in Erewhon is also unrealistic. The protagonists in that story are individualistic adults in strategic search of opportunities to satisfy primarily self-regarding desires.[11] This model is an unrealistic representation of a species in which a prime concern must always have been the protection and nurture of children; a primary characteristic must have been an attachment to family, clan, and tribe; and the crucial factor in sustaining development must have been socially transmitted customs and skills.[12]

The Explanatory Purpose

The point of our unhistorical, unrealistic narrative is to show that despite not having access to prescriptive concepts or practices to begin with, the inhabitants of Erewhon would be more or less inevitably pushed toward the formation of ethical concepts and the development of ethical practices. The idea is not that they would have motives to enter a social contract with one another to establish shared moral standards; even to conceive of such a contract, they would already have to be possessed of ethical concepts. The proposal rather is that, starting as mere reporters, they would be moved in all likelihood to adopt the profile of avowers and pledgers and that, with avowal and pledging established as shared activities, they would be moved in turn to develop properly ethical practices and concepts. The narrative documents an unplanned process of more or less inevitable emergence, not a history of contractual agreement. It is developed

in the spirit of David Hume, who stressed the benefits of an emergence story over any story of a would-be contract.[13]

Even though our narrative focuses on Erewhon, then, it can teach an important lesson about Earth. If it is sound, it demystifies ethics, showing that it can emerge on the basis of the wholly naturalistic elements invoked in the story. Thus it demonstrates that the concepts of desirability and responsibility, and the practices with which they are associated, are not naturalistically mysterious. They are capable of materializing among agents of a kind with human beings and of assuming an important part in the regulation of their lives together. And they are capable of doing this as a result of naturalistically intelligible adjustments to naturalistically plausible opportunities.

Why work with an unhistorical, unrealistic model in seeking to demonstrate the emergability and intelligibility of ethics? One reason is that the model is theoretically tractable. Positing rational agents with defined purposes, determinate abilities, and relatively equal power, it allows us to provide plausible accounts of how they would be likely to respond to certain opportunities, how their aggregate responses would be likely to generate new opportunities, and how they would be likely to respond to these in turn. It enables us to posit and track a more or less inescapable trajectory of development among the inhabitants of Erewhon.

But another reason for working with this model is that the very austerity of its assumptions can help to give us confidence that ethics is inescapable for creatures like us. The individualistic, opportunistic model that it introduces is a worst-case scenario from the point of view of explaining our human fixation on issues of desirability and responsibility. If ethics is inescapable in such a scenario, as the genealogy suggests, then it is even more likely to be inescapable in better-case scenarios. If people would have naturally evolved a sense of ethics in the dry wood of the model, we may hope that they would certainly have done so in the green wood of our actual history.

How does the story presented relate to an historical, evolutionary account of ethics? Histories that purport to tell us about the emergence of ethics often offer only accounts of the emergence of ethical, in particular altruistic, patterns of behavior.[14] What an actual history of ethics ought to provide is a story about the joint, mutually reinforcing emergence of ethical patterns of behavior, on the one side, and of ethical concepts on the other, in particular concepts in the families of desirability and responsibility. It is hard to say how far the conjectural history outlined here

has much to tell us about actual history. But it has at least this positive lesson to teach: that if it is possible to explain how ethics could have emerged under plausible but unhistorical pressures, it ought to be possible to explain how it emerged under the pressures operative in actual history.

THE CONJECTURAL HISTORY OF MONEY

The most familiar analogue to the project undertaken here is the conjectural history of money that is standardly offered in an attempt to demystify financial arrangements: to make sense of money in individualistic, economic terms. The starting state in that story is a barter society—as in our case, a society of relatively equal power—where people are interested in exchanging various commodities or services but, lacking money, cannot easily find suitable partners. You want the dog that I can provide but I do not need the service that you would give me in recompense. I want something that a third person can furnish but that individual does not want my dog or anything else I can currently offer. People in such a society might improve things by writing IOUs in a suitable domain—for example, in the provision of puppies—but this would have similar, if looser limitations. So what might relieve them of the problem they face?

The standard story is that at a certain point it is very likely that some commodity like gold or cattle or tobacco would assume a special status, being recognized as a commodity that everyone wants, or that everyone believes everyone wants, or that everyone believes everyone believes everyone wants, or whatever.[15] And at that point, it would be in the interest of each to gain access to that special good or to IOUs issued by individuals or groups who could provide it. People can be sure of finding providers for the things they want if and only if they have enough of that good, or at least of reliable IOUs in that good, to offer providers an attractive trade.

With these developments, that good and the corresponding IOUs would constitute a medium of exchange, a metric for putting prices on things, and a means of building up purchasing power. In other words, it would become deserving of our name of money. And it would come to resemble our contemporary form of money even more closely if certain other conditions were fulfilled: if the government accepted it in payment of taxes, for example; if the issuers of IOUs became reliable enough to count as banks; if the supply of IOUs was controlled by a central bank that guarded against oversupply and undersupply; and, to mark a recent development in world finance, if those IOUs came to be backed solely by their trading value, not by the guarantee of being able to cash them in.

This narrative demystifies the appearance of money, and our access to the concept of money. It shows that however puzzling it may seem, money is not essentially mysterious: it could have emerged as a by-product of the accumulating, unplanned effects of people's interest in conducting and facilitating trade. The narrative contrasts with a social contract story, for example, because it does not presuppose that people had the concept of money prior to establishing the institution. The idea is that institution and concept would have become simultaneously available in a cascade of individually intelligible developments.

A PHILOSOPHICAL PROJECT

But while the project taken up here is usefully analogized to the economic story about money, it still has a recognizably philosophical character. Think of Wilfred Sellars's myth of Jones, according to which we could have developed concepts of mental experience and attitude, and begun to practice folk psychology, by seeking a theoretical explanation for our dispositions to make certain utterances and to take corresponding actions. Think of David Lewis's demonstration that as self-interested rational agents we could have coordinated with one another in familiar predicaments, and given rise to regularities of the kind exemplified by conventions of language and the like. Think of Donald Davidson's argument that as masters of a finite Tarskian truth theory, we could have become positioned to understand any of a potentially infinite number of sentences. Think of Edward Craig's claim that we could have developed the concept of knowledge, and the practice of justifying claims to knowledge, out of an interest in determining who should count as good informants by criteria available to everyone in the community. Or think of Bernard Williams's explanation of how a community of mutual informants could have evolved norms of truth and truthfulness without relying on any prior sense of a truth-telling obligation.[16]

All of these projects are designed to serve three functions akin to the functions served by the narrative about money. They are meant to identify the putative role and utility of certain practices: the explanatory role of folk psychology; the coordinating role of conventions; the role of recursion in enabling us to understand indefinitely many sentences; the role of knowledge ascriptions in identifying reliable informants; and the role of truth-related norms in organizing a speech community. They are designed to make sense of how people could have come to develop terms and concepts equivalent to our concepts of mental states, social

conventions, sentence meanings, knowledge claims, and truth-related norms. And they do this for each case in a usefully demystifying manner. There is said to be nothing mysteriously first-personal about the psychological understanding to which we lay claim; nothing individualistically unintelligible about our dependence on conventions; nothing impossible for finite minds about understanding indefinitely many sentences; nothing about states of knowledge that makes them more puzzling than other mental states; and nothing about our attachment to truth and truthfulness that requires an independent sense of the obligatory.

What those stories seek to achieve in their respective domains, the story sketched here aims at achieving in the domain of ethics or morality. It seeks, first, to show how practices akin to our ethical practices would be likely to emerge in a purely reportive society like Erewhon; second, to explain why that development would provide referents for the use of ethical concepts like our concepts of desirability and responsibility; and third, to do this in a demystifying way that does not make any naturalistically implausible assumptions.

The approach adopted also resembles the method of creature-construction championed by Paul Grice, foreshadowed by Jonathan Bennett, and used, for various purposes, by philosophers like Michael Bratman and Peter Railton.[17] On that methodology we are invited to imagine how we might design a simple naturalistic creature and build on that design, in successive naturalistic steps, until we come to a creature that can apparently think in familiar psychological and ethical terms. The approach taken here might be recast as an attempt to do something similar at the level of community. The goal is to build on a naturalistic design, in successive naturalistic steps, until we come to a community like the community you and I inhabit where people think in terms of desirability and hold one another responsible for living up to desirable standards.

LECTURE I.
FROM LANGUAGE TO COMMITMENT

BACKGROUND CONCEPTS

This first lecture looks at how the members of the reportive community of Erewhon are very likely to resort to avowals and pledges, and indeed co-avowals and co-pledges, where these do not yet involve them in ethics. The second lecture explores the reasons why the capacity for making such avowals and pledges is going to put them within reach of ethical practices and concepts, leading them to make judgments of desirability and to hold one another responsible to those judgments.

As understood here, reporting, avowing, and pledging are all forms of communication in the sorts of conventional, compositionally constructed signs that are characteristic of natural language. In the normal case of communication, I use those signs with two intentions. The primary intention is to convey some information to an audience and the secondary to achieve that result, at least in part, by making the primary intention manifest to them.[18] Making that intention manifest, by some accounts, involves making it into a matter of common awareness: each of us is in a position to be aware of the intention, in a position to be aware that each is aware of it, and so on.[19] We need not dwell on these complexities here but it is important to recognize that they are in place; they are what distinguish communication in natural language, or so at least it seems, from the transmission of information by the signaling systems used among other species.[20]

Reporting, avowing, and pledging are all varieties of communication in this sense, although they are tailor-made to different domains. I may report any fact about the world or any fact about myself, such as that I have a certain belief or desire or intention. But while I can avow a belief or desire or intention, I cannot avow a fact about the world. And, as we shall see later, while I can pledge an intention, I cannot pledge any other sort of attitude, or of course any fact about the world.

There is a basis for distinguishing between reporting, avowing, and pledging, however, that is independent of the domain in which they may be put to communicative use. And this is the distinction that will be of concern here. It derives from a difference in the extent to which the different speech acts allow me to explain a miscommunication in a face-saving way: that is, in such a way that, if you accept my explanation—if

you think it is credible or adequate—then you will not take me to have been careless or untruthful in the message I conveyed; or at least not as careless or untruthful as I may have seemed. You will not take me to have proved myself an unreliable interlocutor.

The explanation of a failure that deflects the charge of unreliability counts in ordinary parlance as an excuse: it saves my claim to be a cooperative, reliable communicator. Excuses may partially rather than fully explain a failure but for simplicity they will be taken throughout these lectures to constitute full explanations of failure. Reports leave room for two salient sorts of excuses; avowals leave room for just one; and pledges leave room for neither.[21]

Suppose I report to you that something is the case: say, to take a first-person state of affairs, that I weigh less than 170 pounds. And now imagine that you discover that I weigh much more: inviting me to step on an undoubtedly reliable set of scales, it is clear that I am at least 180 pounds. There are two salient sorts of excuses that I may offer in the attempt to show that I was careful about determining the facts and truthful or sincere in communicating them. First, I may offer a misleading-world excuse to the effect that the home set of scales on which I was relying for evidence turns out to be inaccurate. Or second, I may offer a changed-world excuse to the effect that I did weigh less than 170 pounds at the time I made my report, although (sadly) I no longer weigh that now. The misleading-world excuse draws attention to a failure of my words to match the world, the changed-world excuse to a failure of the world to remain matched to my words.

Among the reports I make about the world there are likely to be reports I make about my own attitudes or mind. Thus I may report that I have such and such a belief or other attitude, taking the evidence of introspection or reflection on behavior to show that I believe or desire or intend such and such. And in the case of any misreport on my mind, as with any misreport whatsoever, I may try to excuse it in either of two ways. I may invoke a misleading-mind excuse, arguing that the reason I did not prove to have the attitude I ascribed to myself is that the evidence about what I thought or felt was misleading; I got myself wrong in the way in which I might have gotten a third person wrong. Or I may invoke a changed-mind excuse, claiming that the reason I did not prove to have the attitude—the reason I did not display it later in action—is that my attitude changed before the time for action: on discovering new facts, for example, I ceased to hold the belief I had earlier reported.

The contrast between reports, avowals, and pledges shows up in this domain, where I communicate about my mind. Where I may offer either of two face-saving excuses with an attitudinal report—explanations that aim to save my claim to be a reliable communicator—I may offer only one in the case of an avowal, and neither in the case of a pledge. I take steps in the case of an avowal of attitude that enable me to put aside the misleading-mind excuse and I take steps in the case of a pledge of attitude that enable me to put aside the changed-mind excuse as well. I act in each case so as to deprive myself of the relevant excuse.

Consider the case of avowal first. Here the misleading-mind excuse is unavailable and only the changed-mind excuse can be invoked. Suppose, for example, that I choose to communicate to you that I believe that p, not by reporting on my belief as I might report on the belief of a third party, but just by reporting or asserting that p, thereby expressing my belief state. And now imagine that you discover that I do not actually believe that p: you find that I do not act as if it were the case that p, for example, or you overhear me testifying credibly to a third party that it is not the case that p. How may I excuse my failure to communicate the truth about my belief?

I may certainly claim, with whatever degree of plausibility, that my belief changed since speaking with you, thereby invoking a changed-mind excuse. But I cannot plausibly say that I must have gotten my belief that p wrong when I spoke to you. I showed that I had that belief, after all, by asserting or reporting that p, presumptively in response to the data at my disposal. And knowing that that is so—knowing that this shows that I believed that p—I did not have to consult any introspective or other evidence to determine that I was in that belief state. Thus I foreclosed the possibility of explaining why I misled you by saying later that I was myself misled by such evidence.[22]

Where an avowal rules out one of the excuses that a report tolerates, as in the example given, a pledge rules out both. Suppose I say that I intend to go to your art exhibition this evening and fail to turn up. What I said will count as an avowal of intention insofar as I cannot excuse myself by saying I must have gotten my intention wrong. It will count as a pledge of the intention, however, if it is a matter of common awareness—say, because of the conventions in place—that given how I chose to express myself, I cannot excuse a failure to act on the intention in either of the two salient ways: I cannot claim that I was misled about my mind, in particular my intention, and I cannot say that I changed my mind since

speaking with you. To communicate an intention in this manner, foreclosing both sorts of excuses, is to pledge that attitude as distinct from reporting it or even avowing it.

Avowals and pledges, as conceptualized here, are voluntary acts of communicating an attitude in which I take active steps to put aside one or both of the relevant excuses for failing to display it. The avowed attitude is one that I might have reported, the pledged attitude is one that I might have avowed or reported. In each case it is an attitude about which I might possibly have been misled, as I see things, or which I might possibly change; there is nothing that makes it immune to misreading or alteration. In avowing such an attitude, I set aside a misleading-mind excuse that, by assumption, I might have kept in place. And in pledging an attitude I set aside both a misleading-mind excuse and a changed-mind excuse that, by assumption, I might have kept open.

In describing the developments in Erewhon, charting its transition to ethics, the narrative that follows relies heavily on this excuse-based way of distinguishing between reports, avowals, and pledges. But before beginning to chart those developments it is worth noting two important points. The first is that the excuses introduced in making these distinctions are all epistemic in character and contrast with what we may describe as practical excuses. The second is that the notion of an excuse employed is not itself an ethical notion; it may be understood and employed among Erewhonians, long before they come to ethics.

The excuses invoked in distinguishing between reports, avowals, and pledges all direct us to breakdowns of an epistemic kind. They cite problems that allegedly blocked me from tracking the facts properly, in particular the facts about my mind. In the one case this is a failure on my side to match my words to a misleading mind; in the other it is a failure on the side of the mind to remain unchanged and matched to my words. But there are problems that may be cited in excusing a miscommunication that have a very different, practical character.

Practical excuses, by contrast with epistemic, cite behavioral problems that purport to explain why my assertions or actions did not correspond to what I believed—the assumption is that my beliefs were in order—and why for that very different reason I miscommunicated the facts. They might invoke a problem that inhibits me from telling you what I actually accepted—"I was coerced or induced not to tell the truth"—or a problem that prevents me from acting as I had told you I would act: "I broke a leg before I could do so." Epistemic excuses focus on something that goes

wrong in my processing of information; practical excuses focus on something that breaks the linkage between the information I process and the things I say or do. Both may assume the form of partial excuses rather than excuses of a complete sort, but the assumption throughout these lectures, as noted earlier, is that they come only in the form of complete excuses. This assumption makes the presentation easier, without leaving partial excuses as a mystery; the amendments required by admitting them should be fairly clear in the different cases discussed.[23]

Apart from epistemic or practical excuses, it is useful to recognize a third category of explanation that may be offered on a person's behalf for a failure to live up to their words, whether words uttered in report, avowal, or pledge. This is the explanation that suggests roughly that at the earlier time of utterance or the later time of action the agent was not fully adult or able-minded. The idea is that the agent is exempt, as it is often put, from being held to his or her words, not just excused for failing to live up to them.[24]

It is important to register that the excuses introduced to explain the difference between reporting, avowing, and pledging are of an epistemic character, since otherwise the basis for that taxonomy may seem dubious. But it is even more important to register that, like practical excuses and exemptions, they do not presuppose access to ethical concepts. Otherwise they could not be invoked without circularity in a naturalistic genealogy.

An excuse in an ethical sense would explain an action—say, a miscommunication—in a way that deflects the charge that the agent should be held responsible for acting badly. But excuses as they are invoked here are designed strategically to secure a result that is describable without resort to the notion of responsibility, or any other ethical concepts. They are invoked to show that, despite appearances to the contrary—despite my having uttered misleading words—still I am someone on whom it makes self-interested sense for you and others to rely. Even without access to concepts of desirability and responsibility, there is every reason in Erewhon why I should want to establish such reliability in my own case. And equally there is every reason why I should be able to recognize when others have established it in theirs.

Once we recognize that excuses should be given only a strategic sense in the genealogy provided for ethics, it should be clear that something similar applies to the notions of avowing and pledging. Such acts naturally count, by ordinary criteria, as acts of commitment in which I put my

good name on the line. But they need only be commitments in the strategic, game-theory sense of precommitments, not in any distinctively ethical sense. In making a precommitment, say to performing a certain action, I place a side-bet on doing what I say I will do, where I stand to lose my stake should I fail to do it. By analogy, in making an avowal or pledge, I bet on myself to display the attitude avowed or pledged, where the stake is the cost that I will have to bear if, having failed to display the attitude, I cannot invoke the excuse or excuses foreclosed by the avowal or pledge. That cost will consist in being identified as an unreliable interlocutor.

This lecture is described as moving from language to commitment. But, as these observations underline, the sort of commitment it introduces is strategic rather than ethical in character. The second lecture will move from commitment in that sense to morality proper. It will seek to show how the strategic commitments embedded in the avowals and pledges of Erewhonians will push them towards the adoption of a moral viewpoint, organizing their lives around concepts of desirability and responsibility.

The Reportive Community

Where the genealogy of money begins with a purely barter society, the genealogy of ethics begins with a purely reportive community. This is a community of people, the Erewhonians, who have evolved to the point of being able to use a natural language in communicating with one another but who use it intentionally for the sole purpose of giving reports to one another on how things are in their shared world. I and others in Erewhon make use of conventionally established signs to communicate voluntarily and overtly that things are thus and so: the berries on the hill are ripening, the weather up north is getting better, the prospects for the big-game hunt are looking good. And that is all that we intentionally use such signs to do.

We will each benefit in Erewhon from being able to rely on others to be careful about determining how things stand and to be truthful in making reports; this will expand the range of beliefs on the basis of which we can act with confidence. And we will each benefit by being able to get others to rely on us, going along with the picture we offer of the world and with the plans we make on the basis of that picture. But none of us can expect others to prove suitably reliable or reliant unless we resist the temptation to mislead them and make sure to prove reliable ourselves.

You will have little or no incentive to tell me the truth about what you know, or to rely on me in future, if I have shown myself unwilling to tell you the truth about what I know. On the contrary, you may retaliate against me by refusing some information or collaboration I seek, or by being less than careful in determining the facts I ask about or less than truthful in reporting them to me. You may even retaliate out of an explicit desire to teach me the lesson that if you are to prove reliable and reliant, then I must establish that I too am a reliable person; you may practice a variation of tit for tat in our interaction.[25] In addition, it may also be clear that if I prove to be unreliable you are likely to report this to others—that would help establish your own reliability with them—and so cause me quite a heavy reputational loss.[26]

Under conditions like these it is almost inevitable that we Erewhonians will be generally careful about determining what is the case—this will be supported anyhow by self-regarding motives—and will be truthful in communicating to others what we take to be the case ourselves. We will each generally try to tell the truth, recognizing that this is the only way of establishing a reputation for being reliable and that establishing such a reputation is essential for being able to rely on others or get others to rely on us. The regularity will constitute a social norm or pattern in a more or less familiar sense of the term.

Let a regularity of behavior count as a social norm insofar as conditions such as the following are fulfilled.

- Almost everyone in the community conforms to the regularity.
- Almost everyone expects conformity to attract a good opinion among others and nonconformity—except perhaps in retaliation—a bad opinion.
- Almost everyone is motivated to conform to the regularity, at least in part, by that expectation.[27]

A regularity that fits these conditions is a pattern maintained in a society as a result of the attitudes of the inhabitants toward conformity. That it is actually maintained, as the first condition stipulates, distinguishes it from a standard honored more in the breach than in the observance, such as the ideal of bipartisanship in politics. That it reflects the mutually expected attitudes of inhabitants toward conformity, as the second condition holds, means that it is distinct from a regularity to which others are manifestly indifferent, such as the regularity whereby most people sleep at night, not during the day. And that it is maintained by

that expectation, at least in part, means that it is distinct from a regularity such as taking steps to guard against penury in old age; it is unlikely that people are motivated in any degree by the good opinion they may expect this to win among others.

But while a social norm is distinctive on these three fronts, it may or may not represent a requirement of an ethical or moral kind: that is, a properly "normative" or prescriptive requirement that tells us what to do. A pattern that constitutes a social norm of behavior—say, a pattern of retaliation for injury or discrimination against women—may not be morally permissible, let alone morally required. And a pattern may be morally required, even required by everyone's lights—say, a pattern of moderation in retaliating against injury—without being established as a social norm.

Telling the truth is bound to become a social norm in Erewhon, although in the absence of ethical concepts it will not be marked out by any of us in the society as ethically required or desirable. Each of us will have a strategic interest in winning an opinion and reputation for reportive reliability among others. This will motivate us to be careful about forming true beliefs and, in particular, to be truthful in making reports based on those beliefs. So the upshot ought to be that almost everyone in our society will speak the truth in communicating with others; almost everyone will expect conformity with this regularity to establish a good reputation for them in the minds of others and failures to tell the truth, a bad reputation; and almost everyone will conform on the basis, at least in part, that their reputation depends on it.

For parallel reasons, we may expect our community to establish norms, not just against deception, but also against killing, violence, unfairness, and the like, at least in dealing with other members of the same community. In each of these cases too everyone has a motive, derived from their interest in establishing themselves as reliable partners in interaction, to conform to a suitable regularity, giving rise thereby to a corresponding norm. This observation, as appears in the next lecture, may help to explain why members of the community are likely to converge in developing recognized standards of desirability to which they can hold one another.

Even in the presence of a truth-telling norm, there will be epistemic grounds on which I or someone who takes my side may argue that although I spoke falsely, you should not give up on me as a reliable reporter. I may persuade you that I failed to tell the truth despite taking all the care

I could about determining the facts and despite being truthful in reporting what I took to be the facts. I may be able to show that the world as it presented itself to me was misleading: the berries on which I reported really did look ripe, although perhaps only because of the setting sun. Or I may be able to show that the world changed between the time of my report and your action: a third party came and picked the berries before you got to them. My failing to tell the truth in the presence of a plausible epistemic excuse, whether of the misleading-world or changed-world variety, is not a failure that should induce you to give up on me as a reporter. It explains why I failed to tell the truth in a way that saves my reputation as a truth-teller.

Given that there is a norm of truth-telling in place among us—or indeed a norm of any kind—we Erewhonians can be described as regulating or policing one another into conformity with it. But it is important to distinguish this form of mutual regulation in truth-telling from the practice, which is described in the next lecture, of holding one another responsible—holding one another to account—for telling the truth. The regulation envisaged here falls short of the responsibility practice in two striking ways. First, we may pursue it without being aware of the norm that we regulate one another into sustaining, and so without sustaining it intentionally; and second, we may practice it without any sense of the moral or ethical appeal of the norm.

Within Erewhon, to take up the first feature, we may regulate one another into truth-telling without being aware or conscious of the abstract regularity that we consequently uphold. We may each act in response, now to this individual, now to that, without having any idea of the general pattern we collectively elicit as a result of those responses. Let a rule as distinct from a pattern be expressed by a verbal formula like "tell the truth," which dictates behavior of that patterned kind. While sustaining a regularity or norm of truth-telling, we may not be able to spell out the regularity as a rule; we may not recognize that the upshot of our shared responses on the truth-telling front is to establish conformity with such a rule; and we may not intentionally conform to the rule or intentionally seek to get others to conform.[28]

Our regulation for truth-telling is not only liable to be unconscious and unintentional, however. Whether or not it becomes conscious, to take up a second feature, we are also likely to pursue it without any sense of its moral or ethical appeal. On the story told, we each support the truth-telling regularity or norm out of a desire, now in this case, now in that, to

prove reliable to others. But that is not because proving reliable promises to have impersonal merits of a moral kind. It is only because any failure to prove reliable will make us uncongenial to others and cost us severely; it will involve a strategic loss, diminishing our ability to rely on them in future exchanges or get them to rely on us.

THE AVOWAL OF BELIEF

The Attraction and Accessibility of Avowing Belief

By the story told so far, we Erewhonians are invested in the benefits of mutual reliance and are generally willing therefore to take on the associated costs of proving reliable ourselves; we are willing to be careful and truthful in the reports we make about the world. The investment in mutual reliance means that we must have a particular interest, not just in proving reliable, but in communicating that we desire or intend to be reliable, and that we hold the beliefs that answer to our reports about the world. Thus we must also invest in conveying our attitudes to one another, communicating that we hold this or that belief, are moved by this or that desire, and are bound to this or that intention or plan: presumably, beliefs, desires, and plans that are consistent with peaceful, mutually reliant community. Communicating an account of our attitudes toward one another—in particular, a credible account that ascribes congenial attitudes—we can boost the prospect of establishing relations of mutual confidence and reliance.

What means are we likely to adopt, then, in communicating those attitudes? Should we rely on reporting our attitudes to one another in just the way that we report other aspects of the world? Or should we resort to avowing or pledging? The question arises for our beliefs in the first place, and in the second for other attitudes like desire and intention. The first topic to be discussed, then, is the mode in which we may be expected to communicate our beliefs to one another; the second is the mode in which we may be expected to communicate our other, noncredal attitudes.

The earlier discussion of avowal points up an observation that shows that I will be able to communicate my beliefs to others in Erewhon, not just by making reports on those beliefs, but also by avowing them. Suppose that instead of reporting that I seem to believe that a man I know, Jones, is reliable, I simply say: Jones is reliable. It appears in that case that I cannot excuse my later acting as if Jones were a liar by saying that I must have been misled about the belief I held when I said that he is reliable.

And that being so, the communication I made cannot count as a report. It would allow me to excuse my failing to display the belief expressed by saying that I changed my mind about Jones since the time when I spoke to you. But it would not allow me to invoke a misleading-mind excuse in the same way. It would count by the earlier definition as an avowal of that belief rather than a report.

Why would my saying to you that Jones is reliable communicate, not just that he is reliable, but also that I believe that he is reliable? And why would it communicate this belief in the mode of an avowal that forecloses a misleading-mind excuse?

The key to answering the first question is to recognize the tight link between the idea of making a report and the idea of purporting to have a belief in the content of the report. When I report that Jones is reliable, this linkage means that it ought to be a matter of common belief between us that I purport to believe that he is reliable. The evidence of that linkage is salient for each of us, the evidence that that evidence is salient is itself salient to each of us, and so on.[29] When I make the report that Jones is reliable, then, it ought to be the case that we each believe that I purport to believe that he is reliable, that we each believe that we each believe that I speak with this purport, and so on. Queried about what we believe at any level in that hierarchy we are each going to be disposed, assuming we understand the query, to give the appropriate response.

Thus when I make the report that Jones is reliable I inevitably communicate that I have the corresponding belief. Insofar as the report is intentional, it is intentional on my part that I also convey that information about my belief and that I do so, at least in part, by means of making it manifest that I intentionally convey the information.[30] The semantic message of what I say may be that Jones is reliable but the pragmatic message—the message conveyed by what I do in making the report—is that I *believe* that he is reliable. The semantic message bears on the belief content that I report, the pragmatic message bears on the belief state that I express in giving that report.

Turning now to the second question, why would the pragmatic or expressive message that I convey about my belief in Jones's reliability foreclose my excusing a miscommunication by claiming that the evidence about my mind was misleading? Once again, the key to the answer is the conceptual connection between making a report and purporting to have a belief in its content. This linkage means that I showed that I believe that

Jones is reliable by asserting or reporting, in response to the data at my disposal, that he is reliable. And knowing that this shows that I believed that he is reliable, I can know that I have that belief without consulting any independent evidence about myself, say of an introspective or behavioral kind. Thus I foreclosed the possibility of explaining why I misled you by saying later that I was myself misled by such introspective or behavioral evidence. Not relying on being led by such evidence—and this, as a matter of common awareness—I can hardly claim in excusing a miscommunication to have been misled by it.[31]

Sooner or later we Erewhonians are bound to recognize that communicating a belief by expressing it allows of only one of the excuses that reporting the belief would have permitted. Might we be tempted to resort in that case to play safe and choose to report our ground-level beliefs about any matter rather than communicating them expressively? Rather than expressing the ground-level belief that Jones is reliable, might it push me toward reporting that belief in words such as "My belief seems to be that Jones is reliable," thereby expressing the higher-order belief that I hold the belief that Jones is reliable?

It may at first seem that it would. The cost of my miscommunicating a higher-order belief without being able to invoke a misleading-mind excuse—the cost of miscommunicating a belief about whether I believe that Jones is reliable—would not be very high; you and others are unlikely to rely very much on the truth of a message about my beliefs about my beliefs. But the cost of miscommunicating a ground-level belief about the world without being able to invoke a misleading-mind excuse—the cost of miscommunicating a belief that Jones is reliable—is likely to be quite high; you and others are liable to rely quite heavily on the truth of any such message about my beliefs about the world.

But the resort to higher-order reports about our worldly beliefs would not be likely to attract us, all things considered. For the very fact that the pragmatic or expressive mode of communication allows of only one excuse for error provides a motive for why any one of us should positively cherish it. By communicating a belief in this manner I manifestly take on a greater risk than if I had reported it: I expose myself to the cost of not being able to explain a miscommunication about that belief by recourse to the allegedly misleading character of my mind. And by manifestly taking on such a risk, I make my words more expensive and give you and others firmer ground for expecting them to be true. Why, you may think,

would I take on that risk unless I was pretty sure that I would not have to pay the cost of being unable to excuse a possible miscommunication by appeal to a misleading mind?[32]

Suppose you want to know about my belief about Jones's reliability. And assume that I am anxious to be able to get you to accept whatever communication I make; I am anxious to be treated as someone whose words are credible, both in this instance and more generally. If I hedge and say that it seems to me that I believe that he is reliable, then it will be clear to you that even if I prove not to have that belief, I will be able to get off the hook—I will be able to provide a plausible excuse for having misreported my belief—by saying that I must have gotten my belief wrong. But in that case it will be clear that my words are pretty cheap and are not very credible. If I refuse to hedge in that reportive manner, however, and say simply that Jones is reliable, then it will equally be clear that I have foreclosed access to that easy excuse, that I am taking a considerable risk in communicating my belief in that expressive mode, and that my words therefore are highly credible.

It is bound to be appealing for each of us in Erewhon to give the words we utter as much credibility as we can, assuming we are pretty confident about what we say. And that means that communicating our belief-states pragmatically rather than semantically, at least when they matter in our relationships to others, is bound to be very attractive. Taking on that risk may help to ensure that others actually believe what we say, which is likely to appeal on a number of counts. For one thing, it is likely to get others to rely on us in the instance in question, which may be important for our other purposes. And for another, it will enable us to prove reliable in living up to those words, thereby improving our general reputational standing.

To choose the pragmatic or expressive communication of a belief, foreclosing the misleading-mind excuse but not the changed-mind excuse, is generally to avow that belief, by the terminology adopted here. It is to opt voluntarily for that mode of communication over the salient alternative of just reporting it. And presumptively, it is to opt for that form of conveying the belief, at least in part, because it represents a more expensive and so more credible form of communication; absent that effect, there would be little reason for me or others to opt for avowal. It provides me with a potentially more effective means of getting you to believe what I say.[33]

An avowal will not only be more credible for being more expensive; it will also be more credible for being manifestly adopted on the grounds of

being more expensive. In avowing a belief in conscious awareness of reducing excuses for error, I will manifestly back myself not to have to suffer the cost of being unable, as a result of the avowal, to excuse a failure to act on that belief. I will bet on myself, as a matter of common awareness, not to have to incur that cost. In effect, I will put my money where my mouth is, giving you the firmest grounds for taking me at my word.

While foreclosing resort to the misleading-mind excuse raises the cost and the credibility of an avowal, however, it does not make it prohibitively costly. It leaves the changed-mind excuse in play, for starters, since I will not have foreclosed this. And of course it also leaves room for the various practical, un-foreclosed excuses that I might offer in seeking to establish that despite having miscommunicated a belief, you may take me to be a cooperative and reliable interlocutor. Thus I might excuse myself in the wake of such a miscommunication by explaining that someone had a gun to my head and that I could not speak truthfully without risking my life. And I might achieve the same result by explaining that I was totally disabled from speaking the truth, say as a result of a psychological malaise like paranoia; in that case I am exempted from penalty, in the sense explained earlier. These observations apply to every form of avowal and pledging to be considered in the evolving narrative, although they will not be registered explicitly in every case.

The Feasibility of Avowing Beliefs

These observations suggest that given the ready availability of avowing as distinct from just reporting our beliefs, we in Erewhon are likely to cherish the possibility of avowal; we are likely to rely on avowal to give as much credibility to our communications about our beliefs—certainly to communications that matter in relationships with others—as our confidence allows. But how can we ever be confident enough to put aside the possibility of invoking a misleading-mind excuse for a failure to display a belief avowed? It is one thing for us to have a motive for avowing our beliefs rather than just reporting them. It is quite another for us to be in a position where we have sufficient confidence about what we believe to be able sensibly to take this line.

What might make it possible for me, then, to have the required level of confidence that I know what I believe? What might enable me to have sufficient confidence that I believe that Jones is reliable, for example, or that the berries on the hill are ripening, that the weather up north is improving, or that the prospects for the big-game hunt are bright? I will

recognize any such belief as one that I could be wrong to ascribe to myself, as avowal presupposes; thus it is not like a belief that, as I see it, is true by definition or true on the basis of some unquestioned revelation. So what might make it possible for me to know that I hold it and to be prepared to put aside the misleading-mind excuse?

To answer this question, it is necessary to turn briefly to more general matters. In order to serve a reporting function in a community, natural language must provide the means for speakers like you and me to communicate how things are, according to our beliefs. And this means that when we take care to determine whether or not it is the case that p, to pick an arbitrary sentence from their language, and when we then report truthfully or sincerely that p, we must tend in general to hold the belief that p. If this were not generally the case—if their conscientious reports were correlated only contingently with their beliefs—then their words would be uninterpretable; they would lack the reliable connections with prompting conditions and prompted actions that would enable others to make sense of them.[34]

What is it going to mean, whether in Erewhon or elsewhere, for me to take care about determining that it is or is not the case that p? It cannot mean introspecting my beliefs to see if I actually hold the belief that p. In that case I could never be brought to assent to a proposition that I previously disbelieved or in which I had no belief either way. And taking care over whether or not it is the case that p can often lead me to form a belief in a proposition—that p or that not-p—that I did not previously believe. So the exercise must involve something of a different character.

In the absence of further alternatives, taking care over determining that it is or is not the case that p can only mean attending to the data on whether or not p, where these are mediated by perception or memory or existing beliefs. Attending to those data will lead me to assent to or dissent from the proposition before me, or to withhold judgment. And if I take care to exercise such attention conscientiously, not neglecting any aspect of the data, then I may expect the belief-state to materialize in a suitably robust form. I may expect it to stay in place just so long as the data remain unchanged, and regardless of collateral differences: say, differences in how attractive or unattractive it may be in other respects to hold by that belief.

If I find myself deferring to the data and assenting to the proposition that p after such an exercise, then I can generally assume that I have

thereby come to form the belief that p, whether for the first time or in reaffirmation of a belief already held. In either case I now believe that p in the sense, roughly, that I am disposed robustly to act and adjust as if it were the case that p.[35] If I did not come to form a corresponding belief as a result of such careful assent, then I would not be the sort of being whose language would be interpretable by others.

The upshot of these general considerations appears to be that I can know that I believe something—that is, believe it on a presumptively robust basis—when I find that I assent to it after paying careful attention to relevant data. That conclusion is independently plausible, being borne out by the fact that we often answer the question as to whether we believe that p by thinking about the data and saying sincerely "p"; we often treat it as a question about whether the data elicit in us a belief that p.[36]

But still, the conclusion needs to be qualified, for there is one important complexity to be added to the account of how I can know what I believe. Suppose that I form a belief that p on the basis of data supporting the assertion that p but that there are factors on the horizon of which the two following things are true. First, if they materialized at a certain time, then I would be likely to cease to display the belief formed. And second, I would not be willing to excuse a failure to display that belief by appealing to a change of mind; on the contrary I would be inclined to avow the belief again as soon as those factors went away.

To illustrate the possibility, suppose I am brought by consideration of the data to assent to the proposition that the gambler's fallacy is a fallacy; we may assume that gambling has a place in Erewhon. It may be, first, that the belief is liable to disappear in the excitement of the casino—when there is a run of blacks, I feel sure that red is likely to come up next. And it may be, second, that I would not be disposed in such a case to invoke the excuse that I changed my mind about the matter during my visit to the casino; on the contrary, I would continue to maintain outside the casino that the fallacy is indeed a fallacy.[37]

We may describe any factor that fits these two conditions as a disrupter of the belief I form. It is a disrupter in terms internal to my own practice, not in a sense that invokes external normative standards. What it means for something to count as a disrupter is simply that while it may cause me to drop a belief it affects, it does not induce a change of mind that I would happily cite as an epistemic excuse for no longer displaying the belief. The excitement of the casino would be likely to cause me to

drop the belief that the gambler's fallacy is a fallacy. But even if it did I would not treat the excitement of the casino as an epistemic excuse of the changed-mind variety.

With the notion of a disrupter in place, it is possible to offer a more nuanced account of self-knowledge. As someone competent in the natural language of my Erewhonian community, I can know that I believe something "p"—this is the content of an assertion in that language—when I satisfy two conditions. First, I am evidentially careful in registering relevant data before assenting to it. And second, I am executively careful, if needed, in guarding against potential disrupters of the belief and remaining sensitive to the data.

How can I muster sufficient confidence, then, to be able to avow a belief that p as distinct from merely reporting it? Assuming that I take care to register all relevant data and to avoid potential disrupters—assuming that I take evidential and executive care—the answer is: by deferring to a body of data that robustly elicits it; alternatively, by assenting to "p" or making up my mind that p on the basis of robustly effective data. I know that I believe that p by virtue of knowing what it is that I do in taking that action, whether it be described as deferring to the data, or assenting to the proposition, or making up my mind about it. In a seventeenth-century phrase, I have a maker's knowledge of believing that p, not the knowledge of an observer, even an introspective observer.[38] I can speak for what I believe with an authority of a special, practical sort.[39]

The norm governing truth-telling in Erewhon does not register anything about avowals as distinct from reports. But if the considerations just rehearsed are sound, then it is plausible to expect that in Erewhon we will adjust that norm, consciously or otherwise, to cover avowals. The same is true, not just for norms governing the avowals of belief but also for norms governing the other avowals and the pledges that figure in the discussion that follows. Given the purposes of the narrative, however, it need not register that development explicitly in each case, and need not try to spell out the adjusted shape that the norm would take.

THE AVOWAL OF OTHER ATTITUDES

The Attraction and Accessibility of Avowing Other Attitudes

As it is going to be manifestly attractive for me and others in Erewhon to avow our beliefs rather than just report them, so the same is true for the other attitudes we may wish to communicate to others. Or at least that

will be so with attitudes that are important in our relationships with others and that we must want others to recognize in us. Thus suppose that I am confident in holding about myself that I wish to prove reliable to others, that I prefer talking about a difference to squabbling over it, that I intend to go on a hunt tomorrow, or that I have affection for you as a friend. If I want to convey such an attitude to you with a suitable degree of credibility, then it will be useful to be able to avow the attitude rather than just report it. Avowing it will mean communicating that I have it in such a manner that I cannot excuse a failure to live up to it by claiming that I must have been misled about my attitude. And such a mode of communication will give you much firmer ground for taking me at my word, relying on my possession of the attitude, than if I reported on its presence in a way that kept that excuse open.

But there is a problem in explaining how I can avow a desire or any other noncredal attitude; the means are not so straightforwardly accessible as in the case of belief. I can avow a belief that p, as we know, by asserting that p. But I cannot avow a desire that q by asserting that q; such an assertion would express a belief that q rather than a desire that q. While I may be strongly motivated to avow a desire, then, it seems that I may lack the means of doing so. Certainly I cannot avow a desire in a way that corresponds to the straightforwardly expressive means of avowing a belief.[40]

The expressive means of avowing a belief is of fundamental importance because it is going to be saliently available as well as saliently attractive in Erewhon, even at the stage where we are exclusively interested in making worldly reports to one another. But suppose that the expressive avowal of belief has become standard practice in Erewhon, as the preceding argument suggests that it would. Suppose that it has become a matter of common awareness, in other words, that in Erewhon we will generally want to avow the beliefs we hold rather than just reporting them and that a standard way of doing this is just to express those beliefs: to say "p" in communicating that we believe that p. Under those circumstances, it is plausible that we will begin to recognize other, nonexpressive means of avowing our beliefs. And it turns out that those other means of avowing beliefs offer us models for the avowal of attitudes like desires as well.

The attraction for me and for others in Erewhon of avowing as distinct from reporting beliefs that are important in our relationships with others is going to be obvious to all. And so it is likely to be a matter of common awareness that avowing such beliefs has a much greater appeal

for us than reporting them: we will each have access to evidence of that appeal, of evidence that we each have access to that evidence, and so on.[41] But if this is a matter of common awareness, then the default assumption we will each make with others is that in communicating relevant beliefs to us they are meaning to avow them. They are meaning to speak for what they believe while putting aside the possibility that they may have gotten those beliefs wrong: the possibility that they may have been misled about their own minds.

Suppose then that in the presence of that default assumption, I do not say "p" in expressive mode but resort to the ascriptive mode, as in saying "I believe that p." Should I be taken to be merely reporting on my belief rather than avowing it? It should be clear that in many contexts—specifically, contexts where I act as if I am willing to avow the belief—you will naturally take me to be using the ascriptive remark with the same force that an expressive remark would have had: that is, with the force of an avowal. In such a case you would expect me to go out of my way to indicate that I am merely reporting on the belief, if indeed that was what I was doing. You would expect me to resort to oblique phrasing, as in saying that it seems to me that I believe that p, or that I think that what I want to say is that p, or that I am inclined to believe that p, or whatever.

Absent such phrasing, you will naturally take an ascriptive assertion like "I believe that p"—an assertion in which I ascribe the belief to myself—to have the same avowal force as the expressive assertion "p." And equally you are likely to assign the force of an avowal to other remarks too: say, to an explanatory remark such as "The data explain why I believe that p." In either sort of case, ascriptive or explanatory, you will expect me to be ready to stand by the belief, and not to hedge in the manner of a self-reporter. Hedging in that manner would be unusual enough for you to expect that I would do more to indicate that I was hedging, if indeed that is what I was wanting to do. That this is what would happen in Erewhon is borne out by the fact that this is what happens in actual languages. You would hardly expect to be taken as a mere self-reporter if you said that you believed that Jones was reliable. In order to mark out your utterance as merely the report of a belief you would have to say that your own impression was that he was reliable, or that you were inclined to make the judgment, or something of that markedly tentative kind.

Assuming that this line of argument is sound, consider now the point at which we in Erewhon have established a practice that allows us to avow

our beliefs in ascriptive and explanatory assertions as well as in expressive. At that point, so it turns out, we will have provided ourselves with a salient means of avowing noncredal attitudes as well credal.

Saying in ascriptive mode "I desire prospect R"—say, I desire to prove myself reliable—is not necessarily going to be taken as a mere report that I have that desire but will be heard in appropriate contexts as an avowal. And the same will be true of saying in similar mode that I prefer talking to squabbling, that I intend to go on a hunt tomorrow, or that I like you.[42] Again, saying that there are factors that explain why I desire R or prefer talking to squabbling is not necessarily going to count as a detached explanation but will be taken in suitable contexts as an avowal of the attitude explained. Or at least this will be so with communications in which the attitudes I convey are important in my relationships with others.[43] Thus I will be expected to go out of my way to indicate that I am hedging my bets if that is what I am doing in communicating such an attitude. I will be expected to resort to quaint phrasings, as in saying, "My sense is that I have a desire for R," or "It's possible that I like you," or something of that kind.[44]

The Feasibility of Avowing Other Attitudes

But it is one thing to show that like others I will have a motive and a means of avowing desires and other noncredal attitudes in Erewhon. It is quite another to show that I can be confident enough of having any such attitude to be willing to avow it: to be willing to discount the possibility that I may be misled about my own inclinations. In the case of belief, I can find a sufficient basis for confidence in the fact that the data to which I defer with suitable care robustly elicit assent to the proposition. In order to avow any attitude like desire I need similar grounds to be confident about holding it. But where might I find an effective basis for confidence in this case? The question is particularly challenging because the attitude is one, as avowal presupposes, that I could be wrong to ascribe to myself.

The most plausible answer, which fits with a long tradition of thinking, is that I can find such a basis in the properties that robustly lead me to adopt the attitude, eliciting desire or affection, preference or intention: for short, in the desiderata or attractors present. Thus I can be sure of desiring R insofar as R has properties that attract me to it here and now and that promise to attract me robustly across possibilities where those properties remain in place. Or at least I can be sure of desiring R insofar as I

guard, as needed, against potential disrupters of the desire: that is, factors like wayward whims or impulses that are liable to remove the desire without disposing me to claim an excusing change of mind.

The desiderata that serve to elicit desire come in many different forms. They include neutral properties that can make a scenario attractive for anyone in any situation: that it would be fun, that it would secure peace, that it would reduce suffering. They include agent-relative properties that can make a prospect attractive for anyone in a certain relationship or position: that it would create an advantage for my child or further the prosperity of my tribal group. And they include properties that can make a scenario attractive for anyone with a certain need or taste: that it would satisfy my hunger, relieve my boredom, or preserve my sense of who I am.

By analogy with the case of desire, I can be sure of liking you insofar as your attractive features give you a robust hold on my affections. Or at least that is so to the extent that I also guard against potential disrupters, as I will see them: for example, against the effects of prolonged absence or shifts of mood. Again I can be sure of preferring talking to squabbling insofar as it features desiderata like creating a sense of calm or offering an opportunity for mutual understanding that appeal to me here and now and that promise to remain appealing across a variety of circumstances. Or at least that is so to the extent that I also guard against vicissitudes of taste or inclination that I see as potential disrupters of that preference.

To hold that attitudes of these kinds are grounded robustly in desiderata, as beliefs are grounded in data, is to go along with the idea, long accepted in philosophical tradition, that there are motivating reasons that generally lead human beings to form any such disposition. Those attitudes do not appear out of the blue, so this orthodoxy holds, but are elicited by features that people ascribe to their targets; or at least that is so when they are not subject to disruption and failure.[45]

Although this picture is not endorsed on all sides, it is deeply intuitive. Decision theorists reject it insofar as they treat preferences as primitive rankings, ignoring the possibility that reliable attractors or desiderata lie at their origin. But their view can be seen as a convenient simplification, not a position defended on independent grounds.[46]

Opponents of the picture also include particularists, as they are often known.[47] While they agree that the properties of objects of desire play a characteristic role in eliciting that attitude—or at least in eliciting the corresponding moral judgment—they deny that those properties always weigh in the same direction; the pleasure of an innocent activity may

weigh in its favor, the pleasure of doing something noxious like torturing another may weigh against.[48] But this runs counter to the familiar idea that in deliberating about what we want, we weigh the pros and cons attaching to each option and form our desire on the basis of the resultant effect. And while examples like the pleasure case may seem to put that idea in question, they can be taken equally to show that it is not pleasure as such that counts as a desideratum with us but rather innocent pleasure.[49]

If desiderata play a role of the kind ascribed in this picture, we may expect that we members of Erewhon will have terms for the relevant attractors and will be able to employ those terms to explain our attitudes and by the same stroke avow them. Thus I will explain and avow corresponding desires by describing various scenarios as being a lot of fun or providing a chance to learn something or promising relief from boredom. I will explain and avow a preference for talking over squabbling by pointing to the advantages it offers in generating calm and comprehension. And I will explain and avow an affection by citing your attractive features as a friend. The predication in each case will play the role of an avowal of the relevant attitude insofar as it forecloses a misleading-mind excuse for not acting on that attitude. And it will play the role of an explanation for that avowal insofar as it identifies the property of the object in virtue of which I am robustly drawn to it.

We Erewhonians are disposed to avow rather than report the beliefs we hold, at least when they are important in our relationships with one another: at least when it matters to us that others should rely on our having those beliefs. And by analogy we are disposed to avow rather than just report our desires and other noncredal attitudes, when they matter in our relationships and we want others to rely on our having them. What now appears is that as we can rely on the robust role of data in eliciting beliefs to enable us to avow those states, so we can rely on the robust role of desiderata in eliciting noncredal status in order to be able to practice avowal in this case as well.

On the picture supported, I will form a desire or affection, a preference or intention, insofar as I defer to corresponding desiderata in the way in which I defer to the data supporting a proposition in assenting to it. And I will be in a position to know that I desire or feel, prefer or intend, something—whether for the first time or not—by virtue of knowing that I defer to the relevant desiderata in that way. I make up my mind in response to those attractors and I know the attitudes I form on the basis of knowing what I am doing. It is not by virtue of introspective

observation that I know that I have the attitudes I avow but rather, as in the belief case, by virtue of a sort of maker's knowledge.

THE PLEDGING OF ATTITUDES

By the definitions given earlier, to make a pledge as distinct from an avowal in communicating an attitude is to go one stage further in making the communication credible. It is to raise the cost of the communication by foreclosing not just the possibility of excusing a failure to live up to it by reference to a misleading mind but also the possibility of doing so by reference to a changed mind. If I avow the intention of going with you on a hunt, then I can scarcely excuse my failure to turn up by saying that I was misled about my intention but I can certainly excuse it by saying that I changed my mind since speaking to you. But if I pledge the intention to join you on the hunt, in the sense introduced, then I cannot avail myself of this excuse either. The intention is not immune to being misread, and not immune to change, but the pledge rules out the possibility of my invoking either possibility to excuse my failure to turn up. It will take something like a practical excuse, such as that I broke a leg, to persuade you that I was nonetheless disposed to act in a cooperative manner.

It should be clear that in Erewhon, I and others are going to have a motive for pledging attitudes, if pledging is indeed possible. In particular, we are going to have a motive for pledging the congenial or collaborative intentions that matter in building or maintaining relationships with one another. Pledging an attitude is even more expensive than avowing it, since it exposes me to a greater risk of not having any excuse for failing to act on the attitude. It will be highly credible because of the risk that I choose to take in opting for it. And it will be all the more credible because of the fact that in opting for it I convey the message that I fully recognize the cost of failure but back myself not to incur it.

But however attractive it may be, pledging is only going to emerge in the community if it also proves to be an accessible and feasible option. In order to be an accessible option we will have to be able to identify a linguistic means of communicating a pledge akin to the expressive, ascriptive, and explanatory means of communicating an avowal. But presumably we will be able to find some way of conveying a pledge, if the option of pledging is feasible. And it will be feasible just in case there is some basis, saliently available to each of us, for giving ourselves enough

confidence about maintaining an attitude to be able to pledge it: that is, to be able to rule out not just the possibility that we may have been misled about our attitude but also the possibility that we may yet change that attitude.

The question, then, is whether I could ever have enough confidence in maintaining an attitude to be able to set aside the two standard, epistemic ways of excusing a failure to display it. The question is particularly challenging because the assumption, as we saw, is that the attitude in question is not immune to being misread or immune to change. It arises only with attitudes that, from my perspective, it may actually be wrong to ascribe to myself and it may be wrong to expect to remain unchanged.[50]

Might I be able, then, to pledge a belief? In particular, might I be able to pledge a regular, empirically vulnerable belief of the kind that I recognize I might not maintain; more in a moment on religious beliefs and the like? We live in a changing, incompletely grasped world and although I may think that the data are sufficient to elicit belief in an empirical proposition "p", enabling me to avow that belief in it, I could never be sure that the data would not later be overturned or outweighed. Indeed for me to consider pledging such a belief would betray a misconception about the very attitude of belief. It would show that I did not treat it as responsive to potentially changing data.[51]

Might I be able to pledge any other attitudes besides belief? In order to do so, I would have to be able to identify desiderata or attractors related to those attitudes. And in deferring to those desiderata, I would have to be confident enough about their remaining effective—and about my ability to guard against disruption of their effect—to be able to foreclose the changed-mind excuse as well as the misleading-mind excuse. Is there any reason to think that I might be able to muster such confidence? Surprisingly, there is.

Were I to pledge such an attitude, then the very fact of making the pledge would bring a desideratum or attractor into existence that might serve in the required role. It would make it the case that sticking with the attitude had at least this appealing feature: that it would show that I can be relied upon to keep my word. So the question, then, is whether I could rely on that feature to enable me to pledge a desire for R, a preference for talking over squabbling, or an intention to join you on a hunt. The answer is that I could rely on that feature to be able to pledge an intention but not to be able to pledge any of the other attitudes.

Suppose that I pledge a certain preference—say, for hunting over gathering; that many of the desiderata that attracted me to hunting cease to be appealing; but that I continue to choose hunting because of wanting to show that my word is my bond. Would the preference for hunting remain in place as a result of the pledge? No, it would not. I can hardly count as preferring hunting in the relevant sense—that is, liking it more than gathering—when I only continue to choose it because of having given my word. Preference in the sense at issue here requires me to be attached to hunting on the basis of desiderata other than the attractor that a pledge would put in place. This same sort of problem arises with anything that we are likely to regard as a desire for a prospect R and, of course, with any attitude like affection. I would not count as maintaining the desire or the affection just because I acted as if it were in place but only for the sake of presenting myself as faithful to my word.[52]

This problem does not arise, however, with an intention or plan or anything of that kind. Suppose that in speaking with you I pledge an intention or plan to join you on the hunt, wanting the thrill of chasing prey over open sunny spaces. And imagine that it rains heavily on the appointed day, but that I turn up nevertheless because of having given you my word. Do I count as still holding and acting on the intention pledged? Yes, I do. With an intention as distinct from a desire or affection or preference, the attitude does not have to be sourced in certain sorts of desiderata in order to count as remaining in place. And so the attractor that pledging an attitude creates in favor of maintaining the attitude can serve in this sort of case—although only, it appears, in this sort of case—to give me the confidence required for being able to make a pledge.

We saw earlier that I put myself in a position to avow an attitude on the basis of consciously deferring to a suitable body of data or set of desiderata, where I am careful to register the data or desiderata available and to guard against the possibility of disruption. I know that I think or feel something with sufficient confidence to be able to avow that attitude, by virtue of knowing that I defer to those data or desiderata: by virtue, in that sense, of a sort of maker's knowledge. The same sort of maker's knowledge will enable me to tell that I intend something with sufficient confidence to be able to pledge the intention. In consciously recognizing and deferring to the desideratum that the very act of pledging brings into play—the desideratum that consists in proving that I live up to my word—I can achieve the degree of confidence required. Or at least I can do this to

the extent that I can guard against the disruption of my response to that attractor.

Someone may balk at restricting the speech act of pledging to intentions or plans, on the grounds that many people claim to pledge religious beliefs, political beliefs, and perhaps even beliefs in matters that they take to be a priori: for example, beliefs in classical logic. But the best gloss on such a pledge is to treat it as pledging an intention: say, the intention to treat certain texts or authorities or frameworks as definitive, letting them shape the construal to be given to any other sources of evidence. It is certainly possible to organize life around such voluntarily adopted fixtures. And as this is possible in the actual world, so it would be possible also for those of us who live in Erewhon. The possibility is not relevant to the narrative, however, and will not figure significantly in the evolving story.

The notion of pledging an attitude, in particular an intention, reflects the more regular idea of promising to act in a corresponding way. But the notion of promising in ordinary usage has a strong moral or ethical flavor. It is represented as an act such that if I make a promise to do something, then I have an ethical obligation, however defeasible, to do it. Pledging, as introduced at this point, has no such ethical connotations. When I make a pledge in Erewhon, as when I make an avowal, I back myself to act as thereby advertised, manifestly exposing myself to serious retaliatory and reputational costs in the event of failure. What I do is more akin to making a side-bet that I will hold and act on the intention pledged—a side-bet strategically designed to entice you and others to rely on me—than it is to giving you a promise in the ordinary, moralized sense of that term.

Pledging, by the account offered here, is considerably more costly, and hence more credible, than avowing. If I pledge to act on a certain intention, as in pledging to join you on the hunt, then my stake in living up to those words is higher than my stake would have been, had I merely avowed an intention to join you. And hence you can rely with greater assurance on my joining you than if I had just avowed the intention. But as in the case of avowal, of course, the cost of pledging need not make the act prohibitively expensive. If I fail to join you on the hunt but can invoke the practical, un-foreclosed excuse of a broken leg in explanation of the failure, then I do not lose my stake. And the same is going to be true when I can plead an exempting disability like a temporary bout of insanity to explain the failure. Any such factor can persuade you that, despite the failure, you need not despair of me as a cooperative and reliable interlocutor.

THE CO-AVOWAL AND CO-PLEDGING OF ATTITUDES

The Authorization Presupposed in Co-avowal

When I avow or pledge an attitude I play the role of spokesperson rather than reporter in relation to myself. I speak for myself, as we might say, rather than speaking about myself. I do not convey the attitude in the way in which I might try to communicate the attitude of another, reporting on it in a manner that keeps both epistemic excuses alive. In an avowal I assume the authority to voice an attitude while closing down the possibility that I may have been misled by my own mind. In a pledge I assume the authority to voice an attitude while closing down the possibility both that I was misled about my mind and that I might yet change my mind.

In speaking about your attitudes as a random other, there is a distance between me in the interpreting role and you in the role of the interpreted. In speaking for myself in avowals and pledges—in assuming the role of spokesperson for myself—I reduce or remove that distance. I present myself as the person spoken for and speak, therefore, without fear of an interpretive failure: without fear of a failure to read my mind aright, in the case of an avowal; without fear of misreading or changing my mind in the case of a pledge. Uttered with the authority of a spokesperson, my words are not supported by my skill as the person speaking to track the independently formed attitudes of the person spoken for. They are supported rather by a dual, strategically prudent commitment: as the person spoken for, to conform to what the person speaking says; and as the person speaking, to ascribe only such attitudes as the person spoken for is likely to be willing to display.

As it is possible for me to speak for myself in this way so it is possible in certain contexts for me or someone else to speak for a number of people, being authorized by each of them to make avowals—better, co-avowals—in their collective name, or indeed to make pledges or co-pledges in the collective name. The case of particular relevance to the evolving narrative is co-avowal. Suitably authorized, I will be able to avow a shared attitude in a way that forecloses the possibility of anyone's invoking a misleading-mind excuse—anyone's claiming that I got them wrong—in order to excuse their not displaying the attitude avowed. I will be able to co-avow the attitude in the name of the group.

Co-avowal can mean avowing attitudes in the name of an incorporated agent, as when someone acts as spokesperson for a body like a company or church or state. But when someone co-avows the attitudes of

such an incorporated group they also co-avow the attitudes of its members, qua members. And in that respect the co-avowal is a special case of a more general possibility. The general possibility, which will be the main focus of interest, materializes in any case where I or you or another avows an attitude in the name of the members of a group, whether or not that group constitutes an organized agency. One co-avows the attitude and the rest of us co-accept it: we each treat it as an attitude such that we cannot invoke the misleading-mind excuse for failing to live up to the co-avowal; we cannot claim that in our particular case the spokesperson got the attitude wrong.

It might seem that co-avowal in this sense requires the prior authorization of the spokesperson by other members of the group. That is what Thomas Hobbes assumes when he suggests that the paradigm of authorization is my being appointed by you and others to speak for all of us, as "a representer, or representative, a lieutenant, a vicar, an attorney, a deputy, a procurator, an actor."[53] Hobbes is particularly concerned with the case where I speak for all of us as an organized group agent—say, a corporation or commonwealth—and not just as individuals. But his assumption about the need for prior authorization might be taken to apply to any case of co-avowal, not just to the case where the individuals involved constitute an organized agency.[54]

Advance authorization of the kind at issue may obviously be appropriate in special circumstances where I speak for all of us in a more or less formal capacity. But the authorization on the basis of which I can co-avow certain attitudes in common with you and others need not have its origin in any such ex ante arrangement, however tacit. I may presume on being authorized and claim authorization in the absence of ex post protest at my avowal of a purportedly shared attitude.

On this picture I will signal that I am speaking for what each of us in a certain group thinks or feels, whatever form that signal takes, and I will presume on having the authority of a spokesperson insofar as no one objects to what I say in that role. I do not speak in this case with your advance license, your ex ante authorization. Rather I speak on the presumption that no one will reject my authority and that if no one rejects it, then the absence of rejection will have the same effect as ex ante authorization. You and others do not say "Yea" in advance to my playing the role of spokesperson but neither do you say "Nay" in the wake of my assuming such a role. And that amounts to the same thing. It means that you authorize me in a virtual rather than an actual manner: you authorize me,

not by what you said, but by what you might saliently have said and chose not to say.

It is this general form of presumptive authorization that is of most interest here. It raises two questions, parallel to questions that arose with individual avowal. First, would co-avowal be accessible and attractive for us in Erewhon? And second, would it be feasible? Would any of us have a basis of confidence sufficient to make it into a plausible pursuit: that is, sufficient to give us reason to expect that when we speak for others, they will go along?

The Accessibility and Attraction of Co-avowal

In considering the motives that each of us has for avowing or pledging attitudes, the assumption has been that in communicating with one another in Erewhon we trade independent utterances in a series of exchanges; we each pay the cost of reliably communicating information to another for the reward of being generally able both to rely on others and to induce their reliance on us. On this picture, I make a report or avowal or pledge on my side, you make a report or avowal or pledge on yours. And the main concern on my side, exactly analogous to the main concern on yours, is to prove sufficiently reliable in conveying those messages to be able, as occasion demands, to rely on you and to get you to rely on me.

This assumption about communication is fine for the purposes pursued in the discussion so far. But the accessibility and attraction of co-avowal is going to be obvious only in light of a further observation. This is that in reaping the benefits of mutual reliance, we in Erewhon are bound to pursue exchanges of information that have a more complex, conversational structure. This observation is crucial, because it turns out that conversationally structured exchanges inevitably involve the presumptive form of co-avowal.

Suppose that you and I and others exchange information with a view to resolving a problem we face, whether as individuals or as a community: perhaps a problem about how to resolve a conflict, what to believe about something, or what to do in pursuit of some end. If I am to contribute usefully to a conversation like this, I must speak on the basis of presuppositions about what we each believe and want and intend; if I am wrong about the shared presuppositions then what I say will not engage the concerns of others properly. But in speaking on the basis of such presuppositions I effectively co-avow them in the name of each of us in the conversation. On the presumption that I will not be opposed, it will be manifest to all

that I take the beliefs or desire or intentions as attitudes that we, the members of the relevant group, are each prepared to accept as properly avowed in our joint names.[55]

When the presuppositions are unopposed, the contribution I make to the conversation in expressing a belief or desire or whatever will be to propose, again on the presumption that I am not going to be opposed, that that attitude is also one that we each accept or can be expected to accept as members of the group. If you and others go along with the presuppositions I make and the proposal I put forward, this will establish between us a shared presuppositional base and create a new opportunity for you or someone else to co-avow yet another attitude and, on the presumption that your proposal is accepted, to add further to that base. And if things proceed smoothly along this path, then we may hope to reach a point where our shared set of presuppositions is extensive enough to be able to solve the problem with which we started. It may be enough to eliminate or corral potentially dangerous conflicts, for example, to establish a common belief about some contentious issue, or to make possible the various forms of coordination or incorporation that involve co-pledging.

The presuppositional base built up in such a smoothly progressing conversation is well-described as common ground that we manage to establish between us.[56] It consists in a set of attitudes such that it is a matter of common awareness among those of us engaged in the conversation that we are each prepared to treat those attitudes as properly co-avowed in our name: in that sense, we each co-accept the attitudes. When we go along with a conversation, accepting the different elements in the common ground, we each foreclose the possibility of excusing our failure to live up to the co-avowed attitude by claiming that the co-avower got our attitude wrong.

The attitudes that are built into the common ground between us may include desires as well as beliefs but it is worth noting at this point that there is a great difference between the extent to which the two sorts of attitudes lend themselves to co-avowal. With anything I have solid ground for believing there will be others who share that ground and the belief will be co-avowable in relation to them. But that is not so with all the things I have solid ground for desiring. With some of those things, there may be many people who share that ground but with other things, there may be few or no people who do so. Desires that prove resistant to co-avowal will typically be grounded in agent-relative desiderata to do with what will facilitate my success in some area, help my children, satisfy my curiosity, or whatever.

The difference between belief and desire in these respects will be at the center of concern in the next lecture.

Not all conversations will progress smoothly, of course. Even if my presuppositions are accepted, someone may reject the addition to the common ground that I propose in my initial contribution to the conversation, or indeed in any later contribution. And what goes for me in this regard goes for each of us; none of us can be assured that our contribution at any point will be accepted. But when rejection occurs, this will presumably trigger a round of rejoinders and revisions—it would be in no one's interest just to walk away from every divergence of attitudes—and this can eventually put things back on a progressive path. Conversations in Erewhon may sometimes fail, as they may fail in any society. But, plausibly, they will often succeed.

This image of conversational exchange is easily illustrated. I tell you that there are deer gathering on the southern side of the woods, presupposing for example that we each want to join in a hunt and that we each know where the woods are. You go along with that presupposition, accept my assertion and add, on the basis of the now richer common ground—and perhaps on the basis of the further presupposition that three makes a better hunting party—that a certain friend is available to join us. As the conversation progresses, perhaps now including the friend as well, we each end up co-avowing a desire to hunt. "The hunt is on," one of us may say, or "OK, we're all for hunting." And, explicitly or implicitly, we co-avow a belief that the best time to hunt is now, and we each manifestly avow or indeed pledge a desire and intention to take part, making this too a part of the common ground.

In Erewhon, as in any plausible society, we will each have a motive for taking part in conversations of this kind; after all, they are essential for mutual reliance, enabling us to form, maintain, and develop peaceful, helpful, and collaborative relationships. What the analysis shows is that there is no useful conversation without a pattern of co-avowal and co-acceptance. Contributors each avow attitudes in the name of all those involved, putting them forward as attitudes that everyone avows or can be expected to avow from the standpoint they share. And, whether or not they make any active contribution, participants each accept that any co-avowed attitude that no one opposes is one that they are individually prepared to avow as a member of the group.

This analysis of conversation connects closely with the work of Robert Stalnaker on assertion and related topics.[57] He emphasizes that "the

essential effect of an assertion is to change the presuppositions of the participants in the conversation by adding the content of what is asserted to what is presupposed."[58] And he also recognizes that in presenting certain presuppositions and assertions as expressive of the attitudes of each, every participant presumes on the authorization of others for doing this and is ready to retreat if ex post authorization is denied. Thus he says that the effect of assertion in changing presuppositions, reshaping the common ground between parties, "is avoided only if the assertion is rejected."

What holds about the content of an assertion holds equally, as Stalnaker recognizes,[59] with any presuppositions that an assertion puts in place less obtrusively; an utterance can change common ground, not by just asserting something, but also by intruding a would-be presupposition of all parties. Suppose I say in a conversational context, "The present king of Erewhon is bald." I thereby identify as a would-be presupposition the proposition that Erewhon currently has a king as well as proposing the new presupposition that the king is bald. But you or others can play the same role in rejecting my would-be presupposition as you can in rejecting the content of my assertion. If you each let it pass, then the utterance will count as co-avowing the belief that there is a king of Erewhon, as it will count as co-avowing the belief that the king is bald. You must reject my presumed authority if you are to stop me from changing the common ground in this way.

No man is an island and, as these observations show, no speaker holds just by insulated attitudes. Conversation is essential for gaining the benefits of mutual reliance that we have been emphasizing throughout but it imposes costs on those of us who submit themselves to its discipline. It means that as members of this or that group any one of us may have to avow beliefs and desires in the name of many as well as in our own name alone. And it means that as members of this or that group each of us has to accept that we cannot excuse a failure to live up to any successfully co-avowed attitude by appealing to a misleading mind: that is, by claiming that the spokesperson involved got our mind wrong.[60]

The Feasibility of Co-avowal

So much for the means and motives that I, like everyone else, will have in Erewhon for making co-avowals in the name of others as well as myself. But now, as in earlier cases, we must turn to the question of feasibility. What could give me confidence enough to be ready to speak for a plurality of individuals, avowing a belief or desire as an attitude that you and

others also hold as members of the same group? Where could I find grounds to avow attitudes on the default assumption that they are your attitudes as well as mine?

We saw that in order to avow a belief in my own name I have to think that the data supporting the proposition believed are sufficient to elicit that belief robustly and that my sensitivity to those data is secure: for example, secure against the sort of disruption illustrated by the casino case. And we saw too that in order to avow a desire or other such attitude in my own name, I have to rely on the desiderata at the origin of the desire being sufficient and my sensitivity to those attractors being secure. What might enable me, then, to avow a belief or a desire in the name of you and others as well as myself? Presumably, the fact, as I must take it to be, that you are responsive in the same way to the same data and desiderata, and that you are secure in your responsiveness to them. I must take this to be the case when I venture to speak for us, not just for me, and to avow an attitude in our common name.

Would it be reasonable on my part, or on the part of anyone else in Erewhon, to rely on our being exposed to the same data and desiderata and to be responsive to them in the same manner? It appears to be part of human nature that we exercise joint attention, being consciously directed to matters that we each assume to be available to all, albeit from different perspectives.[61] That being so, I will often be in the position of recognizing that you and others are exposed in common with me to a certain body of data or a certain set of desiderata. Many of us will have access to data not available to others but we may still confront a patently common, intersecting body of data, as when the data in your case suggest that p&q, the data in mine that p&r. Again many of us will recognize special desiderata that are not available to others, and perhaps not available by a sort of necessity: the welfare of your child may matter to you in a way it cannot matter to me, and vice versa. But that is consistent with there being common desiderata or attractors that are effective for both of us: say, that there should be peace and prosperity in the land.

It is one thing, however, to assume that you and I and others may often face a manifestly common body of data or common set of desiderata. It is quite another to suppose that we are each disposed to respond to that common base in a common manner: to suppose that as the data or desiderata are likely to lead me, so in general they are likely to lead you. So the question is whether this further assumption is also a reasonable one to make, in Erewhon or elsewhere.

Suppose that the data you rely on in forming a belief are not good or complete by my lights. Or suppose that the desiderata you are moved by in forming a desire are not attractors that I can see as relevant, even allowing for differences of taste and background, or are only a proper subset of the desiderata I take to be relevant. This will be no problem so long as I can point out my worries about the idiosyncratic or incomplete basis on which you form your attitudes and you respond appropriately. You may change the attitude in response to my complaint or you may show me, perhaps with the help of an anthropologist or psychotherapist, that the basis is not as quirky or patchy as it seemed.

But suppose that you are not disposed or able to do this and continue to display a form of sensitivity to data or desiderata that is completely alien to me. Suppose that without giving me reason to assign different meanings to your words—if it is possible to avoid such differences—you present to me as someone for whom the effect of data is not the same as it is with me; or as someone for whom the role of desiderata is played by different properties from those that make any sense to me. Whether on a wider or narrower front, you and I do not work with the same logic of attitude-formation.

If you were as alien as this, then I could not relate to you as in practice we human beings generally relate to one another. I could only see you as someone to whom I had to adjust, as I might adjust to a force of nature, not as someone conversable: someone I could reach in the space of words.[62] I would be likely to be bewildered and at a loss in such a case. In the end, indeed, I might even be forced to assume that you are a subject for treatment, not conversational interaction.[63]

Elizabeth Anscombe suggests that I would be bewildered and at a loss even if you failed to make sense on quite a narrow front.[64] She argues the point by asking how we would think of a person who seeks something as unlikely as a saucer of mud but cannot do anything to make sense of that desire: cannot present it to us under an aspect with recognizable attractor potential. In order to find the person conversable we would have to see some aspect under which the saucer of mud appeals: say, as an ornament or as a reminder of our mortality. We would not have to be moved by the prospect of having such an ornament or a reminder, as our interlocutor is presumably moved, but we would at least have to recognize why such ornaments or reminders might have an appeal.

Assuming that it is essential for interpersonal interaction that human beings can treat one another as generally conversable, there is no problem

about assuming that in Erewhon we each have a more or less similar sensitivity to data and desiderata, as we do when we co-avow and co-accept certain attitudes. Such a common sensitivity probably comes in good part from our nature. But even if it was not wholly supplied by nature we each would have to simulate or internalize it in order to establish ourselves as someone on whom others could rely and with whom they could converse and do business.

The assumption of mutual conversability is not only needed to explain how we in Erewhon can presume on enjoying a similar sensitivity to data and desiderata, thereby providing a basis for co-avowal. Although not registered earlier, it is also needed to explain why any one of us might be prepared to accept even the individual avowals or pledges of another. You may be content to avow a belief in light of robustly supportive data, and to avow a desire or pledge an intention in light of robustly supportive desiderata. But I would hardly be content to rely on your avowal and pledge if the data or desiderata on which you relied struck me as alien and unmoving. I would be likely to find your avowals and pledges compelling only to the extent that I found that at some level we shared a similar sensitivity to data and desiderata.

As we in Erewhon have a motive to practice co-avowal, then, so we are bound to have the capacity to do so in a range of cases. We can assume a common logic of attitude formation and, identifying the data and desiderata at our common disposal, we can presume on speaking for others as well as ourselves in avowing corresponding beliefs or desires. Or at least we can presume on doing this to the extent that we can assume that all of us who accept a co-avowal made in our name will recognize the possibility of disruption, as we must do in the case of avowals and pledges that we make in our individual names, and will guard against it.

Disruption can be introduced by any factors that might lead us not to live up to a co-avowed or co-accepted attitude without providing us with what we might treat as a changed-mind excuse. The disrupters in the case of co-avowed or co-accepted attitudes will include the sorts of intertemporal factors mentioned in the individual case—the momentary illusion or impulse—as well as disrupters of a particularly interpersonal character such as the tendency to favor a partial perspective or a selfish preference, letting it weaken the force of the data and desiderata shared in common with others. Individuals can expect to be able to make avowals in their own name only if they guard against disruption. And equally they can expect to

be able to co-avow or co-accept attitudes in a common name only if those spoken for guard against the disrupters relevant in this case too.

We saw earlier that it is a sort of maker's knowledge that gives me enough confidence to be able to avow or pledge an attitude in my own name. The line developed here shows that it is something of the same sort that can give me enough confidence to co-avow an attitude in the name of a group. Suppose I recognize that a body of data or desiderata that is available to me in common with you and others is sufficient to elicit a certain belief or desire and that you are disposed to respond to it in the same way as me. In consciously deferring to that base, making up my mind on the attitude to form, I am positioned to know that I hold that attitude. And in consciously recognizing that you and others are responsive in the same way to the same data and desiderata, I am positioned to know that you and others hold that attitude too. Or at least I am positioned to know that you are likely to form that attitude under the stimulus of my co-avowal. I have a comaker's knowledge, as we might say, of our each holding the attitude.

The co-avowals we make may be expressive, ascriptive, or explanatory in form. Thus in the case of the deer-hunt I may say either that the hunt is on or that we're all for hunting or indeed that hunting would be fun or something of that kind. More generally, I may co-avow a belief that p by uttering "p" in a context where this manifestly implicates you and others; or by saying, "We believe that p" or "I take it we all believe that p"; or by claiming that the data show that p or something of that kind. And I may co-avow a desire for R, if not by an expression like "Oh for R," at least by an ascription of the form, "We desire R," or by an explanation such as "R would give us all some pleasure."

This discussion of co-avowal completes a review of the elements needed for the purposes of the second lecture, but it may be useful to add a word on co-pledging as well as co-avowal. Co-pledging an attitude must mean co-pledging an intention, for reasons already rehearsed. It is bound to be attractive for us in Erewhon to be able to pursue common goals with a number of others on the basis of sharing in a co-pledged intention. I will co-pledge an intention on behalf of a group when I speak for what we will do in a manner that closes down the possibility both that I am misled about their intention and that they might yet change their mind. But how could I or anyone else have sufficient confidence that the others in a group are tied in this way to an intention I co-pledge?

One possibility, of course, is that I am authorized in advance to speak in pledging mode for what the group intends to do; or that I am authorized to speak in that way on one aspect of the group performance, others on other aspects; or indeed that a voice constructed out of different inputs from within our membership is authorized as our common voice, whether on all or on only some of the relevant aspects. This, arguably, is what happens when we form a corporate agent, and acquiesce in pursuing the intentions—and the other attitudes—that the authorized spokesperson announces.[65] In this case we act together like a single agent, robustly pursuing a coherent set of goals in robust accord with a coherent set of judgments, where those goals and judgments are determined by the authorized voice.

The other possibility for co-pledging an intention arises when, short of forming a group agent, we the members of a group acquiesce in enacting a particular, episodic goal with sufficient salience for any one of us to be able to presume on the authorization of others in pledging an intention in the name of all the parties involved to realize that goal. This will occur in Erewhon, for example, as it may occur anywhere, when it is salient to all of us in the group—indeed perhaps a matter of common awareness among us—that there is a plan whereby we can achieve something together that we cannot achieve apart; that this is something we each wish to achieve in view of the manifestly present, manifestly effective desiderata; and that if one or more of us takes up their part in implementing the plan, then the others will quickly join in. Under such conditions we members will each go along with the plan and act on a joint intention in pursuing this or that particular end: say, in saving a drowning child or in undertaking a mountain adventure.[66] And in any such case it will be possible for one of us to presume on the authority of others in pledging the intention in the name of all.[67]

CONCLUSION

The main steps covered in the narrative so far should now be fairly clear.

- Erewhon is a society in which by hypothesis we members are competent in natural language but use it only to communicate in reports about our shared world.
- In Erewhon we will want to be able to rely on others and to get others to rely on us and this will give us a motive to prove reliable in giving our reports: to be careful and truthful in what we say.

- That motive will also prompt us to want to communicate our general attitudes to one another, in particular those beliefs, desires, and intentions that help to establish that we are congenial and reliable parties in interaction.
- In communicating our beliefs, one salient option will be to do this by expressing and thereby avowing our beliefs rather than reporting them; to express a belief that p is to assert simply that p.
- Avowal will make our words more costly, since we cannot explain a failure to live up to them, as we might explain words used in a report, by the misleading-mind excuse; and by doing this it will make them more credible and more useful in communication.
- The default attraction of avowal will mean that, unless we go out of our way to indicate otherwise, we will be taken to avow beliefs even when we speak nonexpressively: even when we self-ascribe a belief or explain its hold on us.
- As we will be attracted to avowing beliefs, so we will be attracted to avowing desires and other noncredal attitudes too; we can avow them by means of ascriptive or explanatory modes of self-attribution, if not by expressive ones.
- The attraction of avowal extends also to pledging, which involves closing down not just the misleading-mind excuse but also the changed-mind excuse; pledging an attitude will be more costly, and more credible, than avowing it.
- Pledging may not be possible with beliefs and desires, but it is certainly possible with intentions and plans; pledging an intention will mean setting aside the possibility not just of having misread it, but also of changing it.
- As I may avow beliefs and desires in my own name, so I may co-avow them: that is, avow them in the name of all of us in a certain fixed or fluid group.
- I may do this with the prior authorization of others but also, and more commonly, on the basis of their presumptive authorization; in this case I will count as authorized to the extent that others do not reject what I claim to say in our common name.
- Conversational exchange inevitably involves such presumptive authorization for co-avowal, since it evolves smoothly only when we participants manifestly co-accept the presuppositions and the proposals made by a speaker.
- We are bound to practice co-avowal in Erewhon, since the mutual reliance that we seek is going to materialize fully only to the extent that we can converse with one another and converge on common standpoints.
- This is also true of co-pledging, where that may materialize on the basis either of episodic cooperation between some individuals—acting on a joint intention—or on the basis of incorporating as a group agent.

The social practices that are likely to emerge in Erewhon will have an impact on how we see the world, in particular the social world, we share. They will make various patterns salient to us that would otherwise have been unavailable, so that the view from within those practices will contrast with the view from the bare perspective of natural science. The practices will not direct us to particulars or properties that are naturalistically mysterious, since any patterns that become visible will achieve salience in the process of naturalistically intelligible interactions with one another and with our common environment. But the emerging patterns may still serve to provide us with referents for demonstrably ethical concepts and novel ways of organizing our lives together around those concepts.

This is the guiding promise behind these lectures, holding out the prospect of a genealogy of ethics that serves to vindicate it. The second lecture tries to make good on that promise. It attempts to show how the view from within practices of avowal and co-avowal, pledging and co-pledging, allows notions of desirability and responsibility to gain application and to play an organizing role in the lives of Erewhonians. If successful, the argument demonstrates the near inescapability of ethics for any creatures like us—any creatures endowed with our natural capacities and motives—that have access to natural language. Even if we take ethics at face value as revealing facts about desirability and responsibility, we can insist that it is an unmysterious part of a naturalistic world and that it plays an important role in our lives.

NOTES

1. According to a currently popular version of non-naturalism, pure normative truths are treated like the truths of mathematics and hold with a necessity akin to mathematical necessity. For an excellent presentation, see T. M.Scanlon, *Being Realistic about Reasons* (Oxford: Oxford University Press, 2014). I do not address that position in this text but seek to provide an alternative in light of which I hope it may lose its appeal.
2. Charles L. Stevenson, *Ethics and Language* (New Haven, CT: Yale University Press, 1944); Alfred J. Ayer, *Truth and Logic* (London: Gollanz, 1982); Simon Blackburn, *Spreading the Word* (Oxford: Oxford University Press, 1984); Allan Gibbard, *Wise Choices, Apt Feelings* (Oxford: Oxford University Press, 1990).
3. J. L. Mackie, *Ethics* (Harmondsworth: Penguin, 1977); Richard Joyce, *The Evolution of Morality* (Cambridge, MA: MIT Press, 2006).
4. There are as many forms of non-reductive naturalism, so called, as there are accounts of reduction. I prefer to think that the varieties of naturalism are all reductive and vary only in how constraining they take reduction to be. For a good account of the shape that a reductive naturalism has to take, see Frank

Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press, 1998); and David Chalmers and Frank Jackson, "Conceptual Analysis and Reductive Explanation," *Philosophical Review* 110 (2001): 315–60). And for how it might apply in the ethical case, see also Frank Jackson and Philip Pettit, "Moral Functionalism and Moral Motivation," *Philosophical Quarterly* 45 (1995): 20–40; reprinted in Frank Jackson, Philip Pettit, and Michael Smith, *Mind, Morality and Explanation,* 189–210 (Oxford: Oxford University Press, 2004).

5. Thus the genealogy of ethics, as I pursue it here, is philosophical rather than historical in character, seeking only to tell us how ethics could have emerged, not how it did. And, as noted, it is also vindicatory, not debunking. In both respects, it clashes with the sense of genealogy employed by Friedrich Nietzsche, *On the Genealogy of Morals* (Cambridge: Cambridge University Press, 1997); see Bernard Williams, *Truth and Truthfulness* (Princeton, NJ: Princeton University Press, 2002); and Alexander Prescott-Couch, "Williams and Nietzsche on the Significance of History for Moral Philosophy," *Journal of Nietzsche Studies* 45 (2014): 147–68.
6. Rousseau seems to have thought in this way about his project in the Second Discourse: "The Inquiries that may be pursued regarding this Subject ought not be taken for historical truths, but only for hypothetical and conditional reasonings; better suited to elucidate the Nature of things than to show their genuine origin" (Jean-Jacques Rousseau, *The Discourses and Other Early Political Writings*, edited by Victor Gourevitch [Cambridge: Cambridge University Press, 1997], 132). I am grateful to Alison McQueen for drawing my attention to this. For the various strands in Rousseau's genealogy, see Frederick Neuhouser, *Rousseau's Critique of Inequality: Reconconstructing the Second Discourse* (Cambridge: Cambridge University Press, 2015).
7. I see no essential conflict between the genealogical approach I take and the attempt to seek a naturalistic reduction. Here I differ from Joshua Gert, *Normative Bedrock: Response-Dependence, Rationality, and Reasons* (Oxford: Oxford University Press, 2012), 32–33.
8. Michael Tomasello, *A Natural History of Human Thinking* (Cambridge, MA: Harvard University Press, 2014).
9. The current text takes the notion of reliance that it employs as fairly intuitive. For a useful analysis, see F. M. Alonso, "What Is Reliance?" *Canadian Journal of Philosophy* 44 (2014): 163–83.
10. Christopher Boehm, *Hierarchy in the Forest: The Evolution of Egalitarian Behavior* (Cambridge, MA: Harvard University Press, 1999); and Carles Boix and Frances Rosenbluth, "Bones of Contention: The Political Economy of Height Inequality," *Americal Political Science Review* 108 (2014): 1–22.
11. In assuming that Erewhonians are opportunistic agents of this kind, the genealogy may seem to be of a type with those attempts to show that if we were able to prescribe on the basis of long-term prudence for what we ought to do, then this would support prescriptions of a distinctively altruistic or ethical sort: those prescriptions would present themselves as good prudential policies.

But the genealogy developed here is quite different from any enterprise of that kind, since, for all it requires, Erewhonians at the beginning of the story may not be capable of any prescriptions of the sort that might be based on judgments of desirability—even personal, long-term desirability—or responsibility. The opportunistic rationality displayed in Erewhon need not be mediated, so the assumption goes, by any such prescriptive or normative reasoning. On the relationship between rationality and reasoning, see Philip Pettit, *The Common Mind: An Essay on Psychology, Society and Politics* (New York: Oxford University Press, 1993); and John Broome, *Rationality through Reasoning* (Oxford: Wiley Blackwell, 2013).

12. Kim Sterelny, *The Evolved Apprentice: How Evolution Made Humans Unique* (Cambridge, MA: MIT Press, 2012).
13. David Hume, *Political Essays* (Cambridge: Cambridge University Press, 1994).
14. This charge is laid against a range of studies in Peter DeScioli and Robert Kurzban, "A Solution to the Mysteries of Morality," *Psychological Bulletin* 139 (2013): 477–96, here 478. For an exception, see Philip Kitcher, who describes ethics as "an evolving practice, founded on limited altruistic dispositions that were effectively expanded by activities of rule giving and governance," in *The Ethical Project* (Cambridge, MA: Harvard University Press, 2011), 412.
15. Carl Menger, "On the Origin of Money," *Economic Journal* 2 (1892): 239–55.
16. Wilfred Sellars, *Empiricism and the Philosophy of Mind* (Cambridge, MA: Harvard University Press, 1997); David Lewis, *Convention* (Cambridge, MA: Harvard University Press, 1969); Donald Davidson, *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, 1984); Edward Craig, *Knowledge and the State of Nature* (Oxford: Oxford University Press, 1990); and Williams, *Truth and Truthfulness*.
17. Paul Grice, "Method in Philosophical Psychology," *Proceedings and Addresses of the American Philosophical Association* 68 (1975): 23–53; Jonathan Bennett, *Rationality* (London: Routledge and Kegan Paul, 1964); Michael Bratman, *Shared Agency: A Planning Theory of Acting Together* (Oxford: Oxford University Press, 2014); and Peter Railton, "Reliance, Trust, and Belief," *Inquiry* 57 (2014): 122–50.
18. Dan Sperber and Deirdre Wilson, *Relevance: Communication and Cognition* (Oxford: Blackwell, 1986); and Paul Grice, *Studies in the Ways of Words* (Cambridge, MA: Harvard University Press, 1989).
19. Lewis, *Convention*.
20. Thom Scott-Phillips, *Speaking Our Minds: Why Human Communication Is Different, and How Language Evolved to Make It Special* (London: Palgrave Macmillan, 2015).
21. There is a further excuse that is normally foreclosed even by reporting, or at least by reporting in a shared language: namely, that what I meant by the words used in my report are not what you took them to mean.
22. For an extended, broadly congenial, account of avowals in more or less this sense, see Dorit Bar-on, *Speaking My Mind: Expression and Self-knowledge* (Oxford: Oxford University Press, 2004).

23. Any excuse I offer for a failure to tell the truth, be it epistemic or practical in character, will tend to show that my action was not uncooperative; it was not the product of a lack of carefulness or truthfulness in communication. But as with an excuse for any sort of failure, it may exonerate my action, as it were, without exonerating me as a person. This possibility will materialize in a case where the factor that explains my failure at a certain time—say, ignorance of some fact or an alcoholic hangover—is due to a failure at some earlier time for which I did not have an excuse; in our examples, this will be my not having bothered to learn some important fact or my having drunk to excess. The distinction between excuses that exonerate an enduring agent and excuses that do not have this particular effect is important in other contexts but we may ignore it for our purposes here.
24. On exemptions, see R. J. Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1996) and John Gardner, *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, (Oxford, Oxford University Press, 2007).
25. Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984).
26. Geoffrey Brennan and Philip Pettit, *The Economy of Esteem: An Essay on Civil and Political Society* (Oxford: Oxford University Press, 2004).
27. For an earlier version of this conception of a norm, see Philip Pettit, "*Virtus Normativa*: Rational Choice Perspectives," *Ethics* 100 (1990): 725–55; reprinted in Philip Pettit, *Rules, Reasons, and Norms,* 309–43 (Oxford: Oxford University Press, 2002); and Brennan and Pettit, *Economy of Esteem*. The current version appears in Philip Pettit, "Value-mistaken and Virtue-mistaken Norms," in *Political Legitimization without Morality?* edited by Jörg Kühnelt, 139–56 (New York: Springer, 2008); and Philip Pettit, *The Robust Demands of the Good: Ethics with Attachment, Virtue and Respect* (Oxford: Oxford University Press, 2015). It is modeled on David Lewis's account of convention (Lewis, *Convention*). This notion of a social norm picks up points made in a variety of approaches. See, for example, H. L. A. Hart, *The Concept of Law* (Oxford: Oxford Unviersity Press, 1961); Peter Winch, *The Idea of a Social Science and Its Relation to Philosophy* (London: Routledge, 1963); James Coleman, *Foundations of Social Theory* (Cambridge, MA: Harvard University Press, 1990); Elliott Sober and David Sloan Wilson, *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998); Jon Elster, *Alchemies of the Mind: Rationality and the Emotions* (Cambridge: Cambridge University Press, 1999); and Scott J. Shapiro, *Legality* (Cambridge, MA: Harvard University Press, 2011). For a recent, insightful development of the idea of esteem-based norms, see Kwame Anthony Appiah, *The Honor Code: How Moral Revolutions Happen* (New York: Norton, 2010). And for an overarching theory that is reconcilable with that adopted here, although it uses terminology somewhat differently, see Geoffrey Brennan, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood, *Explaining Norms* (Oxford: Oxford University Press, 2013).

28. Regulation for truth-telling would become conscious in the event, often assumed in discussion of norms, that the three clauses in our definition of a norm were fulfilled as a matter of common awareness among members of the community. In that case we would each see evidence for the pattern of behavior and expectation involved, see that this evidence is available to all, see that the evidence that it is available to all is itself available to all, and so on. In effect, we would each recognize that there is a rule to which everyone conforms on the basis of expectations about how others are likely to respond. While this may be quite a plausible development, it is still worth noting that regulation does not presuppose that it has occurred and may operate without being consciously targeted on upholding a rule.
29. This point is supported at length in Lewis, *Convention*.
30. Standard accounts of communication, as we have seen, require a primary intention to convey certain information and a secondary intention to achieve that result by making the primary intention manifest. But this is unnecessarily strong. It is surely enough, as may hold in the present case, that I intentionally convey the information and intentionally do so by relying on the very manifestness of my intentionally conveying it. To intend a result, in ordinary usage, presupposes that you desire it as such. To bring about a result intentionally requires only that you desire a package that includes the result, not that you desire the result as such. I will communicate my belief that Jones is reliable in saying that he is reliable, even if I do not desire or intend as such to inform you about my belief. It is enough for communicating the belief that I intentionally inform you about it, recognizing that the manifestness of my intentionally doing so will help to bring about that result.
31. There is also a regress argument for this conclusion. The misleading-mind excuse would be in place only if it were the case that I relied upon certain evidence, perhaps of an introspective kind, to determine that I held the belief that Jones is reliable; only in that case could I excuse a miscommunication by saying that the evidence was misleading. But if I had evidence that I hold the belief that Jones is reliable, then that would not only constitute evidence that I do indeed hold the belief; it would also provide evidence sufficient to elicit the higher-order belief that I hold the belief. And presumably I might then express this higher-order belief in words such as: I believe that I believe that Jones is reliable. But if I needed evidence that I believe that Jones is reliable in order to express the belief that he is reliable, then by the same principle I would need evidence that I believe that I believe that Jones is reliable in order to express the higher-order belief. And that would open up an endless regress.
32. On the linkage between expense and credibility in animal signaling, see John Maynard Smith and David Harper, *Animal Signals* (Oxford: Oxford University Press, 2004).
33. Suppose that I report that I believe that p because of wanting to hedge my bets. In that case I will express the higher-order belief that I believe that p. But such an expression may not count as an avowal in the sense defended here: that is, in the sense of an attempt to make my claim more credible. It may be just the

difficulty of going to any higher level, and not the attraction of communicating that belief with maximum credibility, that leads me to express it rather than report it.

34. Davidson, *Inquiries into Truth.*
35. The assumed concept of belief is broadly functional in character, building on the notion of credence in decision theory; see Robert C. Stalnaker, *Inquiry* (Cambridge, MA: MIT Press, 1984) and Philip Pettit, "Practical Belief and Philosophical Theory," *Australasian Journal of Philosophy* 76 (1998): 15–33. There is a serious issue as to how credences relate to acts of assent—and to the states of mind that those acts express—but in this context I generally ignore the problem. See Philip Pettit, "Making up Your Mind," *European Journal of Philosophy* 23 (2015).
36. Gareth Evans, *The Varieties of Reference* (Oxford: Oxford University Press, 1982).
37. Victoria McGeer and Philip Pettit, "The Self-regulating Mind," *Language and Communication* 22 (2002): 281–99.
38. I associate the notion of maker's knowledge with Hobbes and Vico (Philip Pettit, *Made with Words: Hobbes on Language, Mind and Politics* [Princeton, NJ: Princeton University Press, 2008], ch. 1), but Rae Langton cites an employment of the idea in Maimonides (Rae Langton, *Sexual Solipsism: Philosophical Essays in Pornography and Objectification* [Oxford: Oxford University Press, 2009], ch. 13).
39. Victoria McGeer, "Is 'Self-knowledge' an Empirical Problem? Renegotiating the Space of Philosophical Explanation," *Journal of Philosophy* 93 (1996): 483–515; Richard Moran, "Self-Knowledge: Discovery, Resolution, and Undoing," *European Journal of Philosophy* 5 (1997): 141–61; and Victoria McGeer, "The Moral Development of First-Person Authority," *European Journal of Philosophy* 16 (2008): 81–108. See also Alex Byrne, "Transparency, Belief, Intention," *Supplementary Proceedings of the Aristotelian Society* 85 (2011): 201–21.
40. I put aside a possibility like "Oh to q!"
41. Lewis, *Convention.*
42. In the ascriptive avowal of a belief or other attitude, it is worth noting that I do not just communicate that I have the belief—say, the belief that p—or the attitude—say, the desire for R; I also communicate that I have the belief that I hold that belief or harbor that desire. While I give expression to that higher-order belief, however, I do not strictly avow it. This is because it is not for the sake of communicating the belief more credibly, only because I have no option in the matter, that I put aside the misleading-mind excuse in this case.
43. Thus I may not be taken to make an avowal if I communicate an attitude that is manifestly of little or no possible significance for others such as "When I'm alone, I like to read novels."
44. Our observations on the avowal of noncredal attitudes bear indirectly on a familiar debate in metaethics as to what is the relationship between a moral attitude of approval or disapproval and an utterance that communicates the presence of that attitude: say, "I approve of X," or "You ought to do X," or "X

is right." In their simplest forms, one of the standard approaches suggests that this sort of utterance expresses the attitude in the way that an assertion that p expresses a belief, another that it reports the attitude in the way in which an assertion that it seems to me that I believe that p might report a belief. But these standard alternatives—expressivism and subjectivism, as they are sometimes called—are certainly inadequate (Frank Jackson and Philip Pettit, "A Problem for Expressivism," *Analysis* 58 [1998]: 239–51); reprinted in *Mind, Morality and Explanation*, edited by Frank Jackson, Philip Pettit, and Michael Smith, 252–66 (Oxford: Oxford University Press, 2004). In ignoring the role of the belief, simple expressivism would fail to explain why ethical utterances are voluntary acts of communication. In ignoring the difference between reporting and avowing, simple subjectivism would fail to explain why the utterance forecloses the misleading-mind excuse and helps put the speaker on the hook for any failure to live up to the attitude.

45. Philip Pettit and Michael Smith, "Practical Unreason," *Mind* 102 (1993): 53–80; reprinted in *Mind, Morality and Explanation*, edited by Frank Jackson, Philip Pettit, and Michael Smith, 322–53 (Oxford: Oxford University Press, 2004).
46. Philip Pettit, "Decision Theory and Folk Psychology," in *Essays in the Foundations of Decision Theory*, edited by M. Bacharach and S. Hurley, 147–75 (Oxford: Blackwell, 1991); reprinted in Philip Pettit, *Rules, Reasons, and Norms* (Oxford: Oxford University Press, 2002), 192–221; and Franz Dietrich and Christian List, "A Reason-Based Theory of Rational Choice," *Nous* 47 (2013): 104–34.
47. Jonathan Dancy, *Ethics without Principles* (Oxford: Oxford University Press, 2004).
48. Ibid.
49. For a critique of particularism on these general lines—and for a critique of the closely related doctrine I call "interpretivism"—see Pettit, *Robust Demands of the Good*. For a deeper-running complaint about particularism, see Frank Jackson, Philip Pettit, and Michael Smith "Ethical Particularism and Patterns," in *Particularism*, edited by B. Hooker and M. Little, 79–99 (Oxford: Oxford University Press, 1999); reprinted in *Mind, Morality and Explanation*, edited by Frank Jackson, Philip Pettit, and Michael Smith, 211–32 (Oxford: Oxford University Press, 2004).
50. Suppose I have an attitude that I see no possibility, for independent reasons, of ever changing: it might be an idée fixe or an obsessive urge that I cannot seem to drop. With such an attitude I might be able to predict that I will maintain it, and have sufficient confidence to avow that predictive belief. And I might pretend to pledge it, treating it as an attitude that I might have been restricted to reporting or avowing. But I cannot really pledge it in the sense operative here: I cannot take a voluntary step to ensure that I will maintain it. I am indebted to an exchange with Pamela Hieronymi and Jay Wallace on this point.
51. It is possible to be moved to hold a belief by practical considerations, such as the comfort derived from holding it, but it is hardly possible to maintain that

you would stick by the belief for such reasons in face of counterevidence: that it would support your holding by the belief in a suitably robust way.

52. Of course I may pledge to work at maintaining an affection, or perhaps even a desire or preference, committing myself to take steps aimed at preserving the hold of suitable attractors on my sensibility.
53. Thomas Hobbes, *Leviathan*, edited by E. Curley (Indianapolis: Hackett, 1994), ch. 16, 174.
54. Hobbes (ibid., ch. 16) argues that a multitude can become a group agent, being "made one person," by means of advance authorization. He thinks that that is the way that a private body may form—say, a company of merchants, in an example he uses elsewhere—with members authorizing some one officer to make avowals, and indeed pledges, in their collective name under a limited "commission" from them. And he thinks that that is the way in which a commonwealth or state may come into existence, with members authorizing a sovereign spokesperson "without stint." The commission in this case is unlimited, he holds, since he defends an absolutist view of the power that a sovereign has to enjoy if the polity is to be stable and successful. He acknowledges that the entity whose voice a group authorizes may be also a committee that operates by majority voting but denies that it can be a set of mutually constraining individuals or committees such as the competing branches and offices of government that a mixed constitution would allow. He is mistaken on both those counts but this is not the place to explore such issues (Pettit, *Made with Words*; and Christian List and Philip Pettit, *Group Agency: The Possibility, Design and Status of Corporate Agents* [Oxford: Oxford University Press, 2011]).
55. The things I presuppose—or more generally "implicate"—are plausibly going to be identifiable on the assumption that I satisfy constraints like the maxims of conversation—quality, quantity, relation, and manner—analyzed by Paul Grice, "Logic and Conversation," in *Syntax and Semantics*, vol 3., edited by P. Cole and J. L. Morgan, 41–58 (New York: Academic Press, 1975). For a more general perspective, in which relevance is the crucial factor, see Sperber and Wilson, *Relevance*.
56. Robert C. Stalnaker, "Assertion," in Stalnaker, *Context and Content* (Oxford, Oxford University Press, 1999) 78–95; Sperber and Wilson, *Relevance*; and Michael Tomasello, *Origins of Human Communication* (Cambridge, MA: MIT Press, 2008).
57. See also David Lewis, *Philosophical Papers*, vol. 1 (Oxford: Oxford University Press, 1983), ch. 13. For some imaginative applications and developments of the approach shared between Lewis and Stalnaker, see Langton, *Sexual Solipsism*, including the chapter jointly written with Caroline West.
58. Stalnaker, "Assertion," 86.
59. Ibid., 87.
60. Needless to say, the argument here assumes it is acceptable to set aside the effect of power and domination in driving a conversation; this is an aspect of the power equality built into the model, as mentioned in the Introduction.
61. Tomasello, *A Natural History of Human Thinking*.

62. Philip Pettit and Michael Smith, "Freedom in Belief and Desire." *Journal of Philosophy* 93 (1996): 429–49; reprinted in *Mind, Morality and Explanation*, edited by Frank Jackson, Philip Pettit, and Michael Smith, 375–96 (Oxford: Oxford University Press, 2004).
63. P. Strawson, *Freedom and Resentment and Other Essays* (London: Methuen, 1962).
64. G. E. M. Anscombe, *Intention* (Oxford: Blackwell, 1957).
65. List and Pettit, *Group Agency*; and Philip Pettit, "Group Agents Are Not Expressive, Pragmatic or Theoretical Fictions," *Erkenntnis* 79 (2014): 1641–62.
66. There is a large literature on joint intention of this kind. For a congenial perspective, see Bratman, *Shared Agency*, and for other important views, see Raimo Tuomela, *The Importance of Us* (Stanford, CA: Stanford University Press, 1995); Margaret Gilbert, *Joint Commitment: How We Make the Social World* (Oxford, Oxford University Press, 2015); and John Searle, *Making the Social World: The Structure of Human Civilization* (Oxford: Oxford University Press, 2015).
67. The formation of a group agent discussed in the first sort of case is almost certain to involve the members in supporting a joint intention—perhaps voluntarily, perhaps under pressure—to establish an authorized voice behind which they can rally, thereby achieving the coherence of goals and judgments required for corporate agency. See Philip Pettit and David Schweikard, "Joint Action and Group Agency," *Philosophy of the Social Sciences* 36 (2006): 18–39.

LECTURE II.
FROM COMMITMENT TO MORALITY

According to the argument in the first lecture, a simple, reportive community like Erewhon would not be a steady or stationary society. It would contain within itself the seeds of its own transformation, providing its members with motives sufficient to take them beyond giving reports. On pain of having few excuses for failure, those individuals would back themselves to live up to certain self-ascribed attitudes; they would commit themselves in a strategic sense of the term to those attitudes. Their commitments would include avowals and pledges in regard to individual attitudes, and co-avowals and co-pledges in regard to attitudes that they share or expect to share in certain groups.

Nothing in the developments reviewed so far would take the players in our drama into the realm of ethics. They do not make judgments of desirability, and they do not hold one another responsible for living by any judgments of that kind. The challenge in this lecture is to carry forward the project begun in the last and show why the commitments that the protagonists make in avowals and pledges are liable and indeed likely to bring them into ethical space.

The lecture takes up that challenge in two stages: first, by arguing that the players are in a position where it is natural for them to begin to think in terms of desirability; and second, by arguing that having come to think in that mode, they are going to be in a position to hold one another properly responsible to such judgments.

THE NOTION OF DESIRABILITY

Before embarking on the first stage of this argument, it is necessary to articulate what it is to think in terms of desirability. That something is desirable may mean in some contexts that you are permitted to desire it, but it means more generally that you ought to desire it—desiring it is obligatory or mandatory—and this is how it will be taken here. The fact that something is desirable in that sense presupposes that it is one option in a set of alternatives, and requires that you should rank it above the others; it counts not just as desirable in a generic way but, specifically, more desirable than the other options. The alternatives in any such context may be basic alternatives like X, Y and Z or, allowing for ties, disjunctive alternatives like X or Y. In the case of actions these will represent possibilities such that it is up to you whether or not to realize them. Thus you can opt

between doing X or Y or Z or indeed doing X or Y: in this case you choose the disjunction, letting some contingency or chance determine which disjunct is realized.

While the ascription of desirability in the intended sense always introduces a ranking of alternatives, however, it may do so *pro tanto* or *secundum quid*. On the first reading, to say that X is desirable is to hold that it ought to be desired insofar as it displays a certain property or set of properties, F; it ought to be desired qua F, as it is sometimes said. On the second, it means that X ought to be desired *simpliciter;* it ought to be desired outright or unreservedly, not just insofar as it has a certain profile or aspect.[1]

The first task in this narrative is to explain how Erwhonians might develop ranking concepts of desirability, in particular a concept of unreserved or outright desirability. The concept of the unreservedly desirable plays a central role in ethics or morality, so it is assumed here, because of its connection with the more frequently invoked notion of rightness.

As desirability is taken here in a ranking sense—that is, to mean ought to be desired among a presumptive set of alternatives—rightness is taken in a similar way. The question of rightness arises only when there is a set of options in play for an agent or set of agents, and the right option, basic or disjunctive, is that which the agent ought to choose. Is the right option in any such choice set necessarily the unreservedly desirable option? On one pattern of usage, it is: the right option is simply the most desirable option. But on another pattern, the right option is the most desirable option that it would be wrong or blameworthy for the agent not to take. On this second usage, the most desirable option overall will not be the right option if it counts as supererogatory: that is, if it is so demanding that regardless of its desirability, it would not be appropriate to blame the agent for failing to take it.

Should the right option be equated with the most desirable of all the options or with the most desirable option among "erogatory" alternatives? Either equation would work from purposes of the genealogy pursued here, but in what follows rightness will be understood on the second pattern; this has the advantage of registering more clearly the distinction between the obligatory and the supererogatory. On this way of construing the notion of rightness, it is impossible to give an account of how Erewhonians might get to make use of a concept like that of rightness, prior to having an explanation of how they might get to hold one another responsible for how they perform. The concept of rightness can only

appear at the end of the lecture, then, when the issue of responsibility has already been addressed. In the meantime, the focus will be on the concept of desirability, in particular unreserved desirability, and how it might make an appearance in Erewhon.

What is it to think in terms of the desirable? There are three generic and three specific constraints that such thinking must satisfy. The generic constraints apply to thinking in terms of any form of desirability, reserved or unreserved, and indeed to thinking in any prescriptive terms whatsoever: say, as we shall see, to thinking in terms of what is credible or ought to be believed. The specific constraints apply to thinking in terms of unreserved desirability, although perhaps not to thinking in terms of desirability more generally. The generic constraints reflect the role that any judgments of desirability must play in relation to desire, the specific constraints reflect assumptions about the sort of evidence to which such judgments—or at least judgments of unreserved desirability—are responsive.

The first generic constraint is that the desirability of any possible scenario relative to alternatives—say, any option among the options that define a choice—is grounded in the independent features of the alternatives on offer. That scenario cannot cease to be more desirable than competitors without a change in the distribution of independent properties across alternatives; fix those properties and the relative desirability of the alternatives will be fixed too. Why believe in the supervenience of desirability on other properties, as this constraint is often described? The answer is, because it is encoded in the ordinary use of language. When I hold one alternative to be more desirable than another, it is always appropriate to ask about what makes it more desirable: what distinguishes it in independent terms from the other alternatives.

The second generic constraint on desirability judgments is that it is always possible for me or any agent to judge that one alternative in a choice set is desirable, yet not actually desire it; the judgment of desirability can come apart from the appearance of a corresponding desire. This scarcely needs defending, since dissonance and conflict of that sort is a datum of common experience.

The third constraint is that in any such case of divergence, it is going to count as a failure on my part, other things being equal, if I act on my desire and against my judgment of what is desirable. Other things will not be equal, for example, if I make conflicting judgments of reserved desirability, taking one alternative to be desirable under one aspect, a second to be desirable under another, and so on. And other things will not

be equal if my judgment of what is desirable is faulty. But absent those possibilities, the idea is that I will not function properly if I fail to let the judgment of desirability govern what I do. The idea is plausible, since it will be perfectly reasonable to ask me to explain myself in any situation where I fail in that way.

These constraints may be named after what they impose or allow: grounding in the first case; divergence in the second; governance in the third. As they apply to any form of desirability, so they apply to unreserved or outright desirability in particular. The first, grounding constraint, shows up in the fact that if I am told that one option is unreservedly desirable, another not, it always makes sense to ask about what is the difference—the independent difference—between them. The second, divergence constraint, is reflected in our pervasive sense that we may often desire what we think is not unreservedly desirable or fail to desire what we think is. And the third, governance constraint, reflects the fact that we treat judgments of unreserved desirability as having the role of guiding us, and if necessary, correcting us, in the formation of desire and intention.[2]

Apart from these generic constraints, there are three more specific constraints that judgments of unreserved desirability should satisfy; they may also be satisfied by some judgments of reserved desirability but, given the interest in rightness, our focus will be on the unreserved case. These constraints reflect assumptions about the sort of evidence to which judgments of desirability are responsive and may be more controversial than the generic constraints. There are two grounds for endorsing them. First, they fit with plausible, widely supported intuitions. And second, they make the exercise on hand more rather than less difficult to complete: they raise the bar to be crossed in providing a plausible explanation of how residents of Erewhon could come to master and apply the concept of the unreservedly desirable.

The first of the three specific constraints is that when I judge or believe that one among a set of alternatives is unreservedly desirable—when I assent to the proposition ascribing unreserved desirability to it—the property that I ascribe is not the property of being unreservedly desirable$_{me}$, where this is distinct from the property, unreservedly desirable$_{you}$, that you would ascribe if you were the one assenting to the proposition. The constraint is that "unreservedly desirable" is not indexical in the manner of "mine" or indeed "now"; it does not assume a different referent, depending on the identity of the utterer or of the context of utterance.

Thus when I say that it is unreservedly desirable for a person to do something and you deny that it is unreservedly desirable, we are not talking past one another, addressing different properties and responsive to different bodies of evidence. This will be so in either of two salient cases. It will hold if I mean that it is unreservedly desirable for you, independently of position or relationship, to perform the act in question, as in saying that it is desirable in that sense that you relieve pain or promote peace. And equally it will be so if I mean that it is unreservedly desirable for you, given a certain position or relationship, to perform that action, as when I say that it is desirable in that sense for you to favor your child, or if you have made a promise, to keep it.[3]

Where the first constraint holds that judgments of unreserved desirability do not vary in content as between speakers, the second holds that neither do they vary in truth-value. The first constraint is that you and I address the same proposition when, given the same context, I say that something is unreservedly desirable and you deny this. The second is that in such a case at most one of us is correct about that proposition. It cannot be that from my standpoint as an assessor—from the standpoint that my evidence gives me—the alternative at issue truly is unreservedly desirable, and from yours it truly is not; if it is unreservedly desirable from one standpoint, it is unreservedly desirable from all. There may be nothing incoherent about the claim that truth-value may be assessor-sensitive, so that a given proposition should be deemed true from within one standpoint of assessment and false from within another.[4] But, so the second constraint holds, this is not the case with propositions about unreserved desirability.

The third specific constraint on unreserved desirability is that whether an action has this property or not cannot turn on the particular identity of a person, time, or place involved in the action: it cannot be responsive to evidence about such particularities. If it is desirable in that way for someone to do something in this situation, there must be something nonparticular that characterizes that person, and something nonparticular that characterizes that situation, that would make it unreservedly desirable for any relevantly similar agent to perform the action in any relevantly similar situation. This constraint is one of universalizability, as it is often called.[5] It requires that for every particular judgment of unreserved desirability, say that it is desirable for A to do X in situation S, there must be a nonparticular or universal truth to the effect that it is unreservedly desirable for anyone like A—anyone with A's ability, motives, and so on—to perform an action of an X-kind in an S-like situation.

The fact that the concept of the unreservedly desirable satisfies these three more specific constraints implies that the concept of rightness—the concept of the most desirable alternative among 'erogatory' options—satisfies them too. And that implied claim is independently plausible. If the right or the obligatory is to serve its characteristic community-wide role in assessing options and actions, and in determining the responsibility of different agents, then it must be non-indexical and non-relative; it must allow different people to address the same content on the basis of the same criteria of assessment. And equally it must support universalizability by not privileging the particularities of any agent or situation of choice.

Given this understanding of what it is to judge that something is unreservedly desirable, it is possible to explore how far the members of our Erewhonian community, equipped with strategically commissive practices of avowal and pledging, are likely to come to form such judgments. The argument to be offered is that making avowals and co-avowals—in particular, avowals and co-avowals of desire—is going to provide me and others in the community with a perspective from within which it is natural to begin to think in terms of the desirable and the undesirable. Pledges do not figure much in this account but they play a later role in explaining why it is also going to be natural for us to hold one another responsible to standards of unreserved desirability.

THE VIEW FROM WITHIN AVOWAL

When I speak for myself in Erewhon, avowing a belief or a desire—these two attitudes will be the focus of discussion from now on—I rely on a basis for holding the belief or desire that I take to be relatively robust, not just a basis that happens to influence me as a matter of present contingency. The basis for belief is provided by the data at my disposal such that attending to those data, so I take it, elicits the belief. And the basis for desire is provided by the desiderata at my disposal such that attending to them, so I take it, elicits the desire. The data elicit the belief robustly, the desiderata the desire, insofar as the eliciting effect is not a function of some contingent, collateral factor: say, a wish to be someone with such an attitude; absent distorters, it is an effect that data and desiderata may be expected to generate robustly over other variations in my situation.

Given that basis for confidence about the belief or desire, I step out of the contingencies of the here and now when I avow the attitude. Taking

the basis in data or desiderata as sufficient to elicit the attitude robustly, I treat the belief or desire as something I can stand by with relative assurance. I treat it as firmly enough entrenched for me to be able to self-ascribe it in a way that puts misleading-mind excuses beyond my reach.

Or at least I do this to the extent that I take myself to be adequately protected against the disrupting impact of distorters. No matter how effective the protection, I have to recognize that I may occasionally fail to display the attitude ascribed as a result of a distorting influence.

Thus, persuaded by the data to avow that the gambler's fallacy is a fallacy, I still have to recognize that I may lose sight of this truth in the excitement of the casino and that if I do, I will be unwilling to excuse myself by saying that I changed my mind. I will be unwilling to help myself to that excuse, so I foresee, because the change of mind will only have been temporary. Again, persuaded by the desiderata at hand to avow the desire to tell my friends the truth about some embarrassing episode, I may still have to recognize that the shame of doing so face to face may inhibit me from owning up to the episode with some particularly judgmental friends and that if it does, I will be unwilling to say that I changed my mind about wanting to relate the episode to them. I will be unwilling to help myself to that excuse, so I foresee, because the change of mind will have been local to those friends; the desiderata relevant with others will have been present equally in their case.

Think now about how I am likely to view such disrupters, when in the wake of a failure I have to admit that they caused me not to live up to my avowed attitude. I may or may not cite them as practical excuses for the failure, of course, or at least as factors that diminished my practical ability. But whatever I do in that regard, I will certainly disown the actions that they led me to take, whether that be placing a heavy bet on red after a run of blacks, or beating a hasty retreat from meeting with a judgmental friend. I will hold that those actions do not reflect who I am; I will present them as the product of contingent influences or motives that I do not identify with, not as reflections of my robust dispositions.

If I am disposed to take this view in retrospect, however, that has implications for the view I must take in advance of any failure. It means that as I avow the attitude in question, backing myself to live up to it, I must not only hold the attitude avowed and be aware of holding it. I must also assume that I hold the attitude as a result of the impact of relevant data or desiderata, not as a result of a disrupting influence. If I thought that my

holding it was the effect of such an influence, then I would not have the confidence required for avowal.

Thus when I hold an avowed belief that things are thus and so—when I find that scenario avowedly persuasive—I do more than hold by the simple belief that they are thus and so. I hold also by the sophisticated belief that the data support the proposition that things are thus and so, robustly eliciting my belief; or, equivalently, that it is not because of the presence of a contingent distorter that I am led to believe that they are that way. In other words I hold by the simple belief under the assumption—in general, no doubt, a default rather than a confirmed assumption—that there is nothing suspect at its origin. If I thought that there was a suspect distorter at work in eliciting the belief, after all, then that would give me pause about avowing it: I could no longer have the confidence to bet on myself to stick with it.

The same line of thought applies with other attitudes that I avow. When I hold an avowed desire that things be thus and so—when I find that scenario avowedly attractive—I do more than enjoy an attraction to their being that way. I enjoy that attraction but hold at the same time by the belief that relevant desiderata robustly ground the attraction: that the attraction is not due to the contingent influence of any distorter. I stand by the attraction, perhaps letting it shape my actions, under the default or perhaps confirmed assumption that there is nothing suspect at its origin.[6] If I thought that there was a distorting factor at work in generating the attraction then, as in the case of belief, that would give me pause about avowing the desire: I could no longer have the confidence to bet on myself to stick with that desire.

The upshot of this line of argument is that from within the perspective of avowal it is inevitable for me, as it will be inevitable for my fellow Erewhonians, that I should find a use for two presumptively prescriptive concepts: on the one side, that of what I ought to believe, given my avowals of belief; and on the other, that of what I ought to desire, given my avowals of desire. What I ought to believe qua someone who avows beliefs is anything for which, in the presumptive absence of contingent distorters, I find data enough to elicit belief robustly. What I ought to desire qua someone who avows desires is anything in which I find desiderata enough to elicit desire robustly, again in the presumptive absence of distorters. The robustly persuasive, seen from within the practice of avowal, presents as what I ought to believe; the robustly attractive presents from within that practice as what I ought to desire.

Both of these observations are going to be available to me and others in Erewhon and available as a matter of common awareness; the evidence supporting them is salient for all, the evidence that that evidence is salient is itself salient for all, and so on.[7] But that means that not only will each of us in Erewhon be in a position to make use of the concept of what we individually ought to believe and desire—that is, what we ought to believe or desire in light of our avowals; it will also be a concept that we can each use, with manifestly the same referent, in regard to any individual. For each person who avows attitudes we will be able to identify in the one case what we take to be individually credible for that person—this, in the sense of what they ought to believe, not what they may believe—and in the other what is individually desirable for the person.[8]

The concepts of the individually credible and desirable are prescriptive concepts insofar as they satisfy the grounding, divergence, and governance constraints outlined earlier. First, whether something is credible or desirable relative to me is grounded in its relations to data or desiderata; these will explain why a proposition is credible, a prospect desirable. Second, what I find credible may diverge from my actual beliefs, what I find desirable from my actual desires, since distorters may play a role in generating my actual attitudes. And third, assuming that the judgment is not faulty, what I find credible governs or determines what I ought to believe, what I find desirable governs what I ought to desire: this, at any rate, insofar as I go in for the personal avowal of such attitudes. The practice of personally avowing attitudes requires that I ought to believe what I find credible and ought to desire what I find desirable; I could dismiss those requirements as irrelevant only on pain of renouncing the practice.[9] And so it must count as a failure on my part—an inconsistency with what I assume in following that practice—that I do not hold the robustly supported beliefs, or the robustly supported desires, that I avow.

The connection between these prescriptive concepts and the practice of avowal means, in terms introduced above, that the individually desirable is desirable in an aspectual or reserved sense and that something parallel holds of the individually credible. The individually credible or desirable is something I ought to believe or desire insofar as I personally avow the corresponding belief or desire; it is credible or desirable under the aspect it presents from within that practice. And that something is credible or desirable under that aspect does not yet entail that it is unreservedly credible or desirable. The point will be important in later discussion.

With access to the concepts of the individually credible and individually desirable, I and others in Erewhon can form beliefs to the effect that something is credible or desirable in that way. And of course we can even avow such beliefs. Avowal is a potentially recursive operation such that we may avow a belief in a content—that something is credible or desirable in some way—whose very availability to us as a content to be believed itself presupposes the prior use of avowal. While the practice of avowal enables us to gain access to the concepts of the individually credible and desirable, applying these to what we find robustly persuasive and attractive, it enables us at the same time to form and to avow beliefs in propositions that ascribe those very properties of credibility and desirability. This observation applies to all the properties of credibility and desirability to be discussed in this lecture.

How in Erewhon might I avow a belief in the individual credibility of a proposition "p" or in the individual desirability of a prospect R? The usual linguistic devices will be at my disposal. I may express such a belief by saying simply that it is credible that p, or that R is desirable. I may self-ascribe such a belief, and still retain the force of an avowal, by saying that I believe that it is credible that p or that R is desirable. Or I may resort to remarks that serve in context to explain, not why I believe that p or desire R—I may not actually do so—but why it is credible that p or why R is desirable: I may say, for example, "The data stack up in support of 'p' " or "R would be a lot of fun."

These observations show that like others in Erewhon I would naturally be led, just in virtue of making personal avowals, to develop a prescriptive viewpoint on myself. I cannot practice avowal without privileging a robust personal standpoint: the standpoint in which I am responsive to robustly effective data in the case of belief, and to robustly effective desiderata in the case of desire. This standpoint is ideal in the sense that it neutralizes the contingent distorters—the obstacles or limitations—that may affect me as I actually form my attitudes. And so, assuming that standpoint, I can prescribe for how my actual self ought to perform.[10] I can prescribe that actually I ought to stick with a belief that the gambler's fallacy is a fallacy when I go to the casino, or that actually I ought to speak truthfully in face-to-face meetings with my friends. And the wish to live up to my avowals may even lead me to prescribe that should it prove impossible to guard effectively against relevant obstacles or limitations, then I ought to avoid temptation: I ought not to go to the casino or I ought to avoid difficult face-to-face encounters.

The View from within Co-avowal

My individual perspective in avowal lets me identify the robustly and hence avowedly persuasive and attractive, and leads me to give it prescriptive status, treating it as representative of the individually credible, on the one side, the individually desirable on the other. But my perspective in co-avowal, and indeed co-acceptance, allows me to do something parallel at the social level and complicates the prescriptive concepts to which I and others in Erewhon will enjoy access. Before developing this argument, however, it is important to register that co-avowal may be bounded or unbounded and that bounded co-avowal may take as many different forms as there are different bounds.

Co-avowal, Bounded and Unbounded

With any conversation, there is always a projected group of parties to the conversation and there is always a presupposed ground that is accepted in common by those parties. Conversational co-avowal will be bounded if either of these is taken as fixed and allowed to determine the other; it will be unbounded if they are each allowed to change.

There are two sorts of bounded exchange. In a first variety, I and other speakers may seek accommodation with all the members in a given group, being prepared to make compromises—even compromises that disregard what one or another of us sees as relevant data or desiderata—in order to establish common ground. In a second variety, I and other speakers may treat some common ground as so unquestionable—this, perhaps, because each of us takes it to be revealed doctrine—that we are not prepared to give it up for the sake of keeping dissenting members on our side; we are prepared to stand on that ground and hope to find members with whom we can share it. In the one case we keep the members fixed and let the ground move; in the other we keep the ground fixed, at least in part, and let the membership move.

Conversational co-avowal is unbounded when it is not constrained on either front. As a contributor to the conversation I will start from presumptively solid common ground and speak to others who presumptively share that ground; these may constitute a present or just a prospective audience. But in doing this I will remain open to change in two ways: first, by not fixing the membership of the group in advance; and second, by not fixing in advance the ground to be found in common with that membership. I will be happy to let the ground that is co-avowed with

others shift from its initial or any later shape insofar as others change my perception of relevant data or desiderata. Equally I will be happy to let the membership include any others who accept the common ground or are persuasive in arguing for a change, and to exclude any others who reject the common ground but do not provide persuasive arguments in their defense. And as that is true of me, so it is also true of every other participant.

On the unbounded model of conversation, as on the bounded models, I put forward the claims I make on ground that I assume others will share. The others I address in the unbounded case, whether in speech or writing, direct or recorded, are any others who will give me a hearing. I put forward my claims as presuppositions and proposals that I co-avow in the name of such others as well as myself. I essay attitudes on ground that I expect those others—at the limit, perhaps, all presumptively conversable others—to find sufficient. But I am open to the possibility that I may be led by any other to change the ground that I hold fixed and the attitudes I co-avow in our names or co-accept on the basis of another's avowal.

Much of what we say in avowing our standing beliefs and desires, whether in responding to queries, in posting on blogs, in publishing our views, or of course in giving lectures, we say in the spirit of co-avowal. We put forward our attitudes, not in a confessional or autobiographical mode—not on the assumption that our audience is primarily interested in us—but rather in a dialectical mode that invites our interlocutors, real or imagined, to accept what we say or to challenge us where they disagree. As we speak in this mode, we aspire to find a viewpoint that others can share and to contribute to an ongoing conversation. In the unbounded case, we may even think of that conversation continuing into the future or continuing from the past. It was in this spirit that after a day on his farm, the superannuated Machiavelli would enter the courts of ancient men, as he famously records, and feed on the food of their conversation.

Given this distinction between bounded and unbounded conversations and groups, how are things likely to present themselves from within the standpoint of co-avowal? While this issue arises with both bounded and unbounded groups, it assumes a particular importance in the case of the unbounded group and this will be the main focus of attention. Like personal avowal, co-avowal in the name of an unbounded group is inescapable, whereas co-avowal in the name of a bounded group is contingent on happening to belong to such a group. Unbounded co-avowal is inescapable because it is implicit in any exercise of talking things through—and by extension thinking them through—from a standpoint that is

presumptively available to anyone. It reflects interests that none of us can put aside, not just contingent personal interests, or interests contingently shared with a number of others.[11]

From within the Co-avowal of Belief

Suppose that I co-avow a belief that p, opening up a potential, unbounded conversation with anyone at any time or place, however remote. And suppose that some others go along, acquiescing in the co-avowal, offering further co-avowals themselves, and, in an exercise involving various episodes of rejection, rejoinder, and revision, coming to reach a set of beliefs that any one of us is in a position—indeed is manifestly in a position—to avow in all our names and, by aspiration, in the names of any others who join us. In the domain explored this exercise will reveal certain propositions as co-avowedly persuasive: elicited robustly, as a matter of common awareness, by evidence available from within the common standpoint that we share.

It will be manifest to each of us in such a case that due to one or another disrupting factor, we may occasionally fail to believe what is robustly or co-avowedly persuasive within this group: for example, fail to live up to commonly recognized data, as might be illustrated once again in the casino case. But, recognizing what the interpersonally tested data elicit, we must each be disposed to disown any such belief we might form: that is, to treat the factor as disrupting the performance required of us within the standpoint presumptively shared with an open number of others.

This means that what is robustly and co-avowedly persuasive from the common standpoint of this group is a prescriptive category on a par with what is robustly and avowedly persuasive from an individual standpoint. What is robustly and co-avowedly persuasive in this way constitutes the commonly credible, as we in Erewhon might come to articulate it. And the prescriptive status of the commonly credible shows up in its satisfying the grounding, divergence, and governance constraints listed earlier.

What I find commonly credible is grounded in the evidence I identify in common, as I think of it, with an open group of others. It may diverge from what I actually believe under the influence of what I am disposed to see as distorters of that common evidence. And, assuming that my judgment of credibility is not faulty, in such a case it would be a failure on my part not to let it govern my beliefs: it would amount to a breach of the practice in which I rely only on robustly effective data to determine what to believe in common with an unbounded set of others.

How does the commonly credible in this sense relate to the individually credible? The individually credible is that which is robustly elicited, absent distortion, by data I can access on my own. The commonly credible is that which is robustly elicited, absent distortion, by data I can access in common with an open number of others. What counts as data in the one standpoint counts as data in the other; to anticipate later discussion, data are different in that respect from desiderata. But the data available in the common standpoint are likely to be wider than the data available in the individual, so that the common standpoint is bound to have an advantage.

Any data I can access on my own I must treat as accessible in common with others, at least in principle; thus I must be open to co-avowing the belief it elicits, in the name of an unbounded group. But for all I know at any point there may be data accessible in common with others that I have not yet identified; they may only come to light in the future, perhaps only in a future after my death. And so what is commonly credible is going to count as more commanding than what is individually credible. No matter what I previously believed, and no matter what I found individually credible, the discovery that something is commonly credible ought to lead me to believe it henceforth.[12]

The standpoint from within which I believe—and avow the belief—that something is individually credible is idealized, as appeared earlier; it represents a standpoint from within which I can prescribe for my actual self. We now see that the standpoint from within which I believe—and no doubt avow or co-avow the belief—that something is commonly credible transcends and absorbs that standpoint. It represents the ultimate point of idealization from which I can prescribe matters of belief for my actual self. There is no tussle between the individually and the commonly credible, then, and no problem about which to follow in determining what I should hold. This marks a deep contrast, to anticipate later discussion, with the case of desirability.

But as the commonly credible will become defined for the members of an unbounded group, so a counterpart ideal—the jointly credible, as it may be put—is likely to be defined for the members of any bounded group: say, a group devoted to some cause or some creed, whether or not organized as a group agent. How does the commonly credible relate to that which we are liable to find robustly and co-avowedly persuasive from within such a bounded group? How does the commonly credible relate to what is jointly credible, now from within this group, now from within that?

It should be clear that the commonly credible must also transcend the standpoint represented by any such grouping. The beliefs we form in bounded groups are inevitably shaped by a constraint that is independent of data. This may be the desire to find a compromise among a fixed set of members, including perhaps some who are not suitably attentive to the data. Or it may be the desire to stick with a certain core of doctrine, regardless of how far it outruns the data, even perhaps conflicts with the data. Or, to take the case of an incorporated agent, it may be the need to find a set of beliefs that are coherent enough for a group agent to act on, even if this makes it less than fully responsive to the beliefs of members.[13]

The fact that the commonly credible transcends the categories of the individually and jointly credible means that what is commonly credible—what is credible in light of data available in a common viewpoint, open to the future—is going to count as what is unreservedly credible.[14] The commonly credible will be a master category in relation to what is jointly credible in any such grouping, as it will be a master category in relation to what is individually credible for any one of us. This, as will appear, marks a deep contrast between the categories of the credible and the desirable.

From within the Co-avowal of Desire

We now turn from the co-avowal of belief to the co-avowal of desire. Suppose that I co-avow a desire for R, aspiring to speak to an open audience in an unbounded conversation on the topic. And suppose that those who pay attention at any time or place acquiesce in that avowal, offering further co-avowals themselves, and coming in an exercise involving rejection, rejoinder and revision to reach a set of desires that any one of us is in a position—indeed is manifestly in a position—to avow in all our names and, by aspiration, in the names of others whom we allow to join us. Within the domain explored, this exercise will reveal certain scenarios as robustly and co-avowedly attractive for all of us: they will appeal to us in light of desiderata that we are each disposed to acknowledge from within the common standpoint we assume.

What sorts of scenarios are likely to prove robustly and co-avowedly attractive from within this standpoint? The issue is more complex than with the co-avowedly persuasive. What count as data for one are presumably going to count as data for all. But what attracts one person—even what attracts one person robustly, on the basis of recognized desiderata—may fail, even fail with a certain inevitability, to be attractive from a standpoint shared equally with others. I may desire my daughter's welfare

on the basis, precisely, that she is my daughter, where others will only desire her welfare as they might that of a random person. With such an agent-relative desideratum in play, what is robustly attractive from within my individual standpoint may clash with what is robustly attractive from within a standpoint that I purport to share with others. From within my individual standpoint I may avow a desire that my daughter do especially well; from within a common standpoint I may avow a desire that all children thrive equally.

Returning to the question, then, what scenarios are likely to show up as robustly and co-avowedly attractive from within the standpoint of an open group? One set of candidates are those scenarios that are attractive in virtue of promising to realize agent-neutral desiderata we each care about in the same, relatively unconditional way. Plausible unconditional attractors may make it robustly attractive for all of us that norms like truth-telling or non-violence or fair-dealing should obtain; that the species should survive into an indefinite future; that the planet should be able to sustain a high degree of biodiversity; that there should be no unnecessary suffering; and so on.

Another set of candidates for being robustly and co-avowedly attractive for all of us may overlap with this first set. They are scenarios that are attractive to all of us, given that they offer the best prospect of satisfying a certain agent-relative desideratum on the part of each. Suppose we belong to different religions and that it is an agent-relative desideratum for each of us that we should be able to practice our own religion in peace. Even if none of us treats a world with freedom of religion as robustly attractive on an agent-neutral ground about which we converge—even if we do not ascribe a desideratum in common to that scenario—we are likely to treat it as robustly attractive on different, albeit concordant grounds: I, on the ground that in a world of confessional competition it gives me the best prospect of practicing my own religion; you, on the ground that it gives you the best prospect of practicing yours; and so on.

Whether or not they are also robustly attractive on convergent, agent-neutral grounds, there are many scenarios that are likely to be robustly attractive because of enjoying the concordant support of distinct agent-relative desiderata. These will probably include scenarios in which we each look after the welfare of our own children; we each keep our promises to one another; we are each secure against assault by others; and so on. We may or may not find such scenarios suitably attractive on

convergent grounds but we will almost certainly find them suitably attractive on concordant grounds.

We each have to recognize that we may fail to live up to what we find robustly and co-avowedly attractive from a common standpoint due to the influence of potential disrupters. These will include the self-centered preferences that may detach us from the common point of view as well as the wayward impulses that may affect any one of us individually. But when we take something to be commonly attractive, we must assume that we are each going to guard against such disrupters, disowning any desires that they might introduce and seeking to stay faithful to the shared standpoint.

This means that like the co-avowedly persuasive, the co-avowedly attractive is a prescriptive category, directing us to what is desirable from within a standpoint that we share with an open number of others; this is the category of the commonly desirable, as we in Erewhon may think of it. The commonly desirable satisfies the grounding, divergence, and governance constraints associated with all normative or prescriptive categories. It is grounded in the attractor properties, convergent or concordant, that make something attractive to me and others qua members of an open group. It may come apart from what I actually desire—say, as a result of disruption—even when I purport to occupy a common standpoint with others and to think as a member of an open-ended group. And assuming that it is not faulty, I should allow my judgment of the commonly desirable to guide or dictate what I actually desire, at least so far as I operate as one member in an open group of others. Since the commonly desirable is responsive to the robust attractors that our practice in co-avowal and co-acceptance takes as determinants, it would be a failure on my part, so far as I genuinely share in that practice, to endorse some conflicting desire of my own.

As avowal is recursive, so too is co-avowal. Once the category of the commonly desirable becomes available in Erewhon, we members are likely to form, avow, and co-avow beliefs in propositions to the effect that this or that scenario is commonly desirable. And this, despite the fact that the property of being commonly desirable will only have become available to us in virtue of our having practiced co-avowal in respect of what we desire.

As I may avow a belief in the individual desirability of a scenario by expressive, ascriptive, and explanatory devices, so I may resort to such

devices in co-avowing a belief in its common desirability. Depending on context, I can co-avow the common desirability of a prospect, R, by saying that it is desirable or commonly desirable, by saying that we desire it or believe that it is desirable or commonly desirable, or by explaining its desirability appropriately: for example, by reference to how much fun it would be or to how it would give us each fair returns.

I am naturally led, just by virtue of making individual avowals of desire, to develop a prescriptive point of view: a robust personal standpoint from which I can judge my actual performance, letting what I desire be assessed in terms of whether it is individually desirable. In the same way I am naturally led, just in virtue of the co-avowal and co-acceptance of desire that I practice in unbounded conversation, to develop a second prescriptive standpoint on desire: a robust common standpoint from which I can judge what I desire, letting it be assessed in terms of whether it is commonly desirable.

How does the individually desirable relate to the commonly desirable? In many cases they may coincide, as when it is both individually and commonly desirable that I should tell the truth to others; it is something I would prescribe for my actual self both from the robust personal standpoint and the robust common standpoint. But it should be clear that in many cases these standpoints are quite likely to come apart. The role of agent-relative desiderata in determining what is individually desirable means that what I would prescribe from an ideal, individualized standpoint may diverge from what I would prescribe from the ideal, socialized counterpart.

This means that neither the individually desirable nor the commonly desirable can play the role of the unreservedly desirable. The categories target what is robustly productive of desire, in the one case under the identification of robustness that goes with my practice of individual avowal, and in the other under the identification that goes with our practice of common co-avowal. Those practices may come apart in a way in which the corresponding practices in the case of belief do not. And so neither has a position in relation to the other that might give it a claim to direct us toward a master category: that of the unreservedly desirable.

We have been looking at what is likely to count as co-avowedly attractive from the point of view of an unbounded group and at what is commonly desirable in the sense of being robustly attractive from within the standpoint of the group. But the argument developed in the case of that group suggests that we can derive parallel conclusions for this or that

bounded group. As the perspective of the unbounded group will direct us to the category of the commonly desirable, so the perspective of any bounded group will point us toward the category of what is jointly desirable for members of that group in their part as members; this will be identified by what proves to be robustly attractive to them in that role.

We the members of Erewhon, like the members of any society, are likely to find ourselves in any of a number of bounded groups; indeed our own community, as distinct from neighboring societies, will constitute one example. And within such a partial grouping, as within the unbounded community imagined, we will each conduct conversational exchanges with others in which we co-avow and co-accept a range of desires that reflects the properties that matter from our shared standpoint, identifying scenarios that we will see as jointly attractive.

These properties will include group-relative properties that matter to us as members—the welfare of our club, the prosperity of our community—but that may not matter to us in other roles. And so for each such grouping there is likely to be a notion of the jointly desirable that operates prescriptively but is in potential conflict with rival forms of desirability. It is liable to conflict with what counts as jointly desirable from the standpoint of other bounded groups. And of course it is liable to conflict with what counts as individually or commonly desirable.

The upshot of these observations is that each of us in Erewhon is going to be led by the inexorable pull of avowal and co-avowal into countenancing a range of prescriptive, idealized standpoints. As I avow a desire in my own name, I have to privilege what is robustly I-attractive, treating the possibility of disruption as a failure against which the avowal requires me to guard. And as I avow a desire in the name of a group, I have to privilege the robustly we-attractive in a parallel way, where that may mean what robustly attracts us in an unbounded group or what robustly attracts us in one or another bounded grouping.

The category of the robustly attractive that I identify in each case corresponds to a distinct prescriptive ideal of desirability, satisfying the constraints outlined earlier. What I see as desirable in the individual, common, or joint sense must be grounded in the properties that serve as robust attractors from within the corresponding practice. What I see as desirable in any such sense may diverge from what I actually desire. And, on pain of breaching the requirements of the relevant practice of avowal or co-avowal, I must let the perception of what is desirable in any of these senses govern the desires I actually form in the relevant area. In each case,

fidelity to the practice will require me to let the desire that is avowed or co-avowed under it have a governing role in determining what I actually desire.

TOWARD THE UNRESERVEDLY DESIRABLE

A Breakthrough and a Setback

If the argument so far is sound, then in the wake of the developments charted, I and you and others in Erewhon will enjoy a conceptual breakthrough but suffer at the same time a conceptual setback. The conceptual breakthrough occurs in the areas of both belief and desire, the setback is confined to the area of desire alone.

The breakthrough is that we will become able to think in prescriptive terms, enjoying a position from within which we can distinguish between things as we actually believe or desire them to be and things as we ought to believe or desire them to be. How we ought to believe and desire things to be, in this way of conceiving of them, is how we would hold or want them to be, if we conformed to the constraints associated with a standpoint we privilege. Depending on context, this is the standpoint of the avowed self, or the self projected in one or another form of co-avowal. And by parallel it may be the standpoint of the pledged self, or the self projected in some form of co-pledging. It is the standpoint of the self as spokesperson for itself, now in one context, now in another.

It is a real gain for us in Erewhon to be enabled on this basis to think and talk in prescriptive terms, avowing beliefs as to what is credible and we ought to believe, what is desirable and we ought to desire. In the purely reportive life we enjoyed previously, we might have responded to incentives now in this manner, now in that; we might have generated in aggregate social patterns like those of general truth-telling; and we might even have been in a position to recognize and welcome that result. But there would have been no standpoint from which we could have seen ourselves as measuring or not measuring up to one or another ideal. And there would have been no basis for personal aspiration and criticism, or aspiration and criticism interpersonally shared.

All of that changes with the development of acts of avowal and pledging, and the appearance of the concepts of credibility and desirability that they bring onstream. Those shifts enable us to recognize that how we are may or may not be how we are committed by relevant practices to being, and that when we do not conform to the requirements of those practices then we display a sort of failure. We fall short in ourselves of the self

we spoke for; we believe what is not credible, or desire what is not desirable, by the lights of that bespoken self. And when we recognize the actuality of failure, we simultaneously grasp the possibility and attainability of success. We see it as within our grasp: what we can become, if only we let the bespoken self shape the self we actually are.

The perspective of the bespoken self is also, it should be noted, the perspective of the beholden self. For the self we speak for in avowing or pledging, co-avowing or co-pledging, is a self that we have given others the right, under the rules of relevant practices, to expect us to display. It is a self such that if we do not display it, then the rules of avowal or pledging give them the right to ignore certain excuses, to treat us as uncooperative, and to impose associated retaliatory and reputational costs.

This conceptual breakthrough ought to be welcome in itself, opening up a wholly new way of thinking, and holding out the possibility of aspiration and criticism. But it ought also to be welcome insofar as it is bound to serve our interest in being able to rely on others and to get others to rely on us. For with the extra resources available in any given context, we will each have an enhanced capacity to assure others of our reliability. I will be able not just to avow or co-avow a belief that p or a desire for R, but to avow or co-avow a belief that p is credible or that R is desirable. And in reaching for such an extra means of communicating my belief or desire, inviting you to rely on me, I will signal that I must pay an even heavier reputational cost, should I fail in the absence of excuse to live up to what I say.

But while the breakthrough into prescriptive space is a huge benefit for us Erewhonians, it comes in the area of desire at a serious cost. Although it may serve us well in this or that insulated context—say, in a context where just individual desirability, common desirability, or one or another form of joint desirability is relevant—it will not do so when there is a prescriptive clash between those modes of desirability. Those modes of desirability are all aspectual in character, as we have seen, and different options may be desirable in different modes; one of my alternatives may be individually desirable, another commonly desirable, and yet another desirable from the joint standpoint of some contingent grouping. It might be, for example, that in a time of need it is individually desirable that I devote my efforts to my children, jointly desirable from the standpoint of my neighborhood that I devote them to the welfare of those who live nearby, and commonly desirable that I put them at the service of people as a whole: say, the general population of Erewhon.

I must be able to resolve this problem in any particular case, deciding which option I should take in light of the rival claims, since otherwise I will be irredeemably ambivalent, unable to decide between the conflicting standpoints. And I must be able to resolve this problem in a way that is manifest to my fellows, since otherwise I will be unable to present myself as a non-ambivalent agent that they understand and on whom they can rely. I will be multiply and inconsistently bespoken, on the one side; multiply and inconsistently beholden, on the other.

BEYOND THE SETBACK

I might resolve this problem brutely by declaring in each particular case, or in cases generally, that such and such a mode of desirability is the winner, without providing any explanation of why it scores better than the rivals. That would not be an appealing way to go, however. It would be tantamount to letting a lottery decide the issue and would project the image of being a more or less random self, not a self for which I can speak and expect to command a hearing: not a self that I can expect to be taken seriously by others, or indeed by myself. It might enable me to escape ambivalence but would do so only at the cost of embracing arbitrariness.

But there is a more appealing way for me or any other to resolve the problem raised. This would be to treat the grounding attractors that determine the individual or common or joint desirability of options as features that may be weighed against each other across categories, determining in aggregate which option in any given choice is to be selected. The local balance of features in one practice determines what is individually desirable from my perspective, in another what is commonly desirable, in yet another what is jointly desirable from the standpoint of this or that grouping. The idea in this resolution would be to let the global balance of features across those different categories determine what is desirable according to me in a practice-neutral sense: in a sense that treats no particular practice as special.

It may be that we in Erewhon depend on practices of avowal and coavowal to access concepts of the individually, commonly, and jointly desirable. And it may be that none of those concepts has the status of a master category, as the concept of the commonly credible has that status in the case of belief. But with any choice we face that still leaves us with the salient possibility of allowing the desiderata that support competing, practice-relative judgments of desirability to enter into competition with one another and to determine which alternative in the choice answers

best to the desiderata as a whole. It leaves us with the option of recruiting the desiderata mobilized within each practice in a further role, letting them interact with one another to fix what counts as desirable in a sense that is no longer tied to any particular practice.

Given the notions of what is desirable in one or another practice-relative manner, the concept of what is desirable in a manner that is not bound to any particular practice ought to be readily available to us. And it ought to be clear that the introduction of such a concept would serve an important function in our psychologies. It would enable me—or you or any other—to escape the specter of ambivalence. And it would do this without exposing me to the charge of arbitrariness in how I make up my mind.

The concept of the practice-neutrally desirable ought to be attractive in Erewhon, not just because of enabling each of us to resolve intrapersonal conflicts, but also because it holds out the prospect of making certain interpersonal conflicts resoluble as well. I and you and others will identify the practice-neutrally desirable on the basis of the aggregate desiderata—recognized as a matter of common awareness between us—that stack up in support of different options in any choice. And that means that there is at least the possibility that we may be able to agree about the option the agent should choose—perhaps even co-avowing or co-accepting a belief in its desirability, whether across a bounded or unbounded community—and that we should conceive of the issue as one that we may sensibly debate; we should conceive of it as an issue shared across people, not as a different issue for each person.

We will certainly be able to achieve agreement in any case where one option in the choice satisfies all the desiderata satisfied by others, and satisfies them in a higher measure or, satisfying them in at least equal measure, satisfies other desiderata as well. And equally we will be able to achieve agreement in other cases to the extent to which the weightings we attach to different desiderata are in more or less the same range. The desiderata that we recognize may be weighted differently to the point where there is no agreed resolution available in certain cases as to what it is desirable for the agent to choose, in which case it will be indeterminate whether this or that option is practice-neutrally desirable. But the concept of practice-neutral desirability will at least allow us to think that we may achieve resolution and not have to give up in advance on the prospect; we may put ourselves in a position to co-avow a belief in the desirability of this or that alternative in a choice.

The role that the concept of the practice-neutrally desirable can play in helping to resolve interpersonal as well as intrapersonal conflicts should lead us in Erewhon to cast it as a concept capable in principle of resolving both. Construed in that way, the concept would hold out the prospect of a result that we ought to embrace, given our interest in establishing relationships of mutual reliance. Those relationships will be the more readily available, the more we can converge with one another in our judgments of practice-neutral desirability.

These observations argue that as we in Erewhon would each come to access a range of practice-relative concepts of desirability, so in all likelihood we would evolve a corresponding concept of practice-neutral desirability. This argument marks a crucial development in the narrative, for the concept of the practice-neutrally desirable is vanishingly close to the concept of the unreservedly desirable, as was outlined earlier. The concept of the practice-neutrally desirable would not single out the options to which it applies under certain aspects only; it would go beyond what is desirable only pro tanto—only under the aspect it presents from within a certain practice—to what is desirable simpliciter. And, so it turns out, it would satisfy both the generic constraints and the specific constraints that the concept of the unreservedly desirable may be expected to satisfy.

The notion of the practice-neutrally desirable straightforwardly satisfies the generic constraints of grounding, divergence, and governance, If one option in a choice counts as practice-neutrally desirable and others not, then there must be a difference in the desiderata that ground their relative desirability. If an option is desirable in that sense by my judgment, it may still be that what I desire diverges from that judgment; the desiderata supporting the judgment may fail as a result of disruption to elicit the desire. And, assuming that the judgment is not faulty, it ought to dictate or govern what I actually desire. It would be a manifest failure on my part not to let the judgment play that role, for it would amount to a failure not to let my desire be guided by the desiderata registered in the judgment.

The first of the specific constraints on the concept of the unreservedly desirable requires that you and I should have the same content in mind when we judge that it is desirable for anyone, whether anyone in general or anyone in a certain position, to choose a given option. And the second requires that it should be true or false that the option is desirable—assuming the issue is determinate—ruling out the possibility that it might be true by your criteria as an assessor, false by mine. The concept of the

practice-neutrally desirable is bound to meet these constraints insofar as it is designed to facilitate the resolution of interpersonal as well as intrapersonal resolution. If it is to play that role in any range of cases, it must rule out both the relativity of content that the first constraint forbids and the relativity of truth-value that the second constraint outlaws. It must direct us to a range of issues that we may hope to explore and perhaps resolve in common.

Finally, does the notion of the practice-neutrally desirable satisfy universalizability, the third more specific constraint associated with the unreservedly desirable? Is it the case, for example, that when I say it is practice-neutrally desirable for someone independently of position or relationship to relieve suffering, I must hold that it is desirable in the same sense for anyone to relieve suffering? And is it the case that when I say it is desirable in that sense for someone in the position of a parent to give special care to their own child, I must hold that it is desirable in the same sense for anyone in a parental relationship to give such care to their own child?

The plausible answer in each case is, yes. What makes an action practice-neutrally desirable under the story told about Erewhon is the fact that in aggregate the desiderata derived from relevant practices weigh up in favor of the action. But whether they are agent-neutral or agent-relative in character, the desiderata must be general properties, if they are to enable each of us to prove conversable to others. In the agent-neutral case the general property might be that of relieving suffering, in the agent-relative that of being an agent who looks after their own child. And so, if they weigh up in favor of A's doing an action X in situation S, then they must weigh up in favor of the relevantly similar B doing an X-like action in any S-like situation.

These considerations argue that the concept of practice-neutral desirability, which Erewhonians would be likely to evolve under the pressures described, can be identified with the familiar concept of unreserved desirability. With this argument then, it becomes plausible to endorse the claim that the practices of avowal and pledging that are likely to emerge in Erewhon would push inhabitants to come to think, not just in terms of practice-relative forms of desirability, but in terms of an outright form as well: in terms of desirability, period.

The concept of what is right or obligatory figures more prominently in ethics than that of the unreservedly desirable. But, as noted earlier, if Erewhonians come in addition to have the ideas of responsibility and blame, they are also going to be in a position to introduce a concept that

plays the role of the right, under the construal adopted here. It will be right for someone to choose a certain option under that construal if it is unreservedly desirable that they should choose it and if they would be blameworthy—if they would be fit to be held responsible—for not doing so. And that means that the concept of the right cannot be introduced into the narrative until it has been extended to encompass responsibility as well as desirability.

Some Observations about Unreserved Desirability

Before moving on to issues of responsibility, however, it is worth making some observations about the Erewhonian concept of practice-neutral desirability. These will help to display the implications of identifying that concept with the concept of the unreservedly desirable.

It is likely, to make a first point, that we in Erewhon will agree about the practice-neutral desirability of many types of choice, or at least about the desirability of most instances of those types. Consider choices of the kind that are generally resolved by strategically supported social norms of the kind considered in the first lecture. Each of us is likely to be sensitive to the desiderata, relevant to individual, joint, or common desirability, that argue for the practice-neutral desirability of conforming to such a norm. Thus it is likely that we will agree in thinking that at least when other things are equal, it is practice-neutrally desirable to tell the truth, abstain from violence, not steal what belongs to others under local conventions, and the like. It is likely that with such examples we would be prepared to co-avow or co-accept a belief in their desirability, whether on a bounded or unbounded basis.

Although we may readily agree on the extension of practice-neutral desirability in such run-of-the-mill cases, however, a second point to note is that we may disagree strongly in other cases. Indeed we may even think that there is no fact of the matter to be resolved in those cases: we may treat the question as to whether a certain choice is practice-neutrally desirable as indeterminate.

That we are likely to disagree about the desirability of various choices derives from the fact that we may differ in the relative weights that we assign to relevant desiderata—nothing in the narrative rules out this possibility—and be led by those desiderata in different directions. While disagreeing about such choices, we may each think that our opponents are wrong and that further reflection on the desiderata—say, on the importance they give those desiderata in other contexts—would lead them

in our direction. Or, despairing of even the theoretical possibility of reaching agreement, we may conclude that there is no resoluble fact of the matter at issue between us: the question dividing us is indeterminate.

Moved by the costs to the mother, for example, I may think that abortion is practice-neutrally desirable, at least in certain cases; moved by the prospects for the foetus, you may think that taking the child to term in such cases is practice-neutrally desirable. And confronted with that divide, we may treat the difference as one of a disagreement that is worthy of further discussion. Or we may decide that the issue is indeterminate, reflecting an indeterminacy about the relative importance of the costs to the mother and the prospects for the foetus.

But this observation about possibilities of disagreement and indeterminacy should be balanced against a third point: that in any such case, there is always a prospect of conceptual evolution and eventual convergence. Mutual conversation and exchange may reveal that a desideratum we take to support a judgment of desirability in one context applies in a context where we hadn't invoked it previously and requires in consistency that we make a corresponding judgment of desirability there.

Thus it may be that invoking the notion of equality in arguing against discrimination between males, we may be led to recognize that it also argues against discrimination across gender. We may come to a ground-level agreement on the extension of the property of practice-neutral desirability to such a case. Or, consistently with ground-level disagreement, we may at least agree at a higher level that the extension to that case is determinate. In the case of higher-level agreement we may co-avow or co-accept the desiderata, suitably weighted, that we take to make the question resoluble. In the case of ground-level disagreement, we may also co-avow or co-accept certain particular judgments of desirability.

A fourth and final observation about practice-neutral desirability bears on the issue of amoralism. We are likely to react negatively to anyone's failure to agree with us, at least when the evidence is clear, about something that we regard as determinately desirable: that is desirable in terms of commonly endorsed, commonly weighted desiderata; for short, desirable in terms of accepted standards. But will this be the case, even if the person claims to be an amoralist who does not recognize the category of the practice-neutrally desirable? Yes, it will. Amoralists are likely to have a hard time of it in Erewhon.

Amoralists can scarcely reject the claims of different modes of desirability, since these appear in light of more or less inescapable practices.

And they can scarcely deny the relevance of the desiderata invoked in those practices, since that would put their very conversability in question. How then can they deny the possibility of allowing these desiderata, in the event of conflict—or at least in the event of some conflicts—to determine in aggregate the option that is practice-neutrally desirable? Certainly they cannot deny this without argument. And since they will come out as losers in most arguments, at least by the views of their adversaries, they are likely to be treated like self-serving offenders.[15] Thus they will not be allowed any excuse for refusing to act as it is practice-neutrally desirable for them to act, by the common perceptions of the community.

THE CHALLENGE OF RESPONSIBILITY

The observations made in this discussion give us solid ground for thinking that Erewhonians would evolve a conception of desirability akin to that with which people operate in more familiar worlds. They would inevitably develop practices of avowing and pledging, co-avowing and co-pledging their attitudes, as registered in the first lecture. Those practices make it more or less inevitable that they would introduce practice-relative notions of desirability. And, confronted with conflicts between those notions, it is equally inevitable that they would develop the concept of what is practice-neutrally desirable: in effect, so it was argued, the concept of what is unreservedly desirable.

If it is to give us a potential explanation of the emergence of ethics, however, the narrative must also explain how the inhabitants of Erewhon can come to think in terms of responsibility as well as desirability. It must explain how we who have evolved the concept of the unreservedly desirable would go on to hold one another responsible in various choices for not selecting the unreservedly desirable option. That is the next challenge that the genealogy has to confront.

The Notion of Responsibility

As in the case of desirability, it is essential in pushing forward this project to have a good sense of what fitness to be held responsible connotes in everyday usage and practice; otherwise it will not be clear what is needed for the narrative to achieve success. There are various accounts in the literature of what it means to hold someone responsible for having done something. But rather than going into the debate between these approaches, the line taken here will be to present an account that has two now familiar considerations in its favor. First, it satisfies many of the

common connotations of saying that someone is fit to be held responsible for an action. And, second, it offers a rich account of those connotations that makes the task to be discharged by the narrative about Erewhon more rather than less difficult to accomplish; it does not tilt the scales in favor of success.

What responsibility connotes in ordinary usage is best articulated for the scenario where I hold you responsible for something I see as an undesirable choice: an offence or misdeed. This is a case in which the implications of being fit to be held responsible are sharp and the costs high, so that the received understanding of responsibility is likely to be at its clearest. And if it proves possible to articulate the concept of responsibility for this scenario, then the lessons should carry over to the case where I hold you responsible for having done something good rather than something bad.

Suppose, then, that I hold you responsible for a misdeed of some kind. Let this be an action like telling a lie, when there are no special considerations that make it desirable in the context on hand to hide the truth. On the account to be adopted here, there are three aspects to holding you responsible in this way or, alternatively, to treating you as fit to be held responsible. They come out nicely in three distinct messages that I convey if in the case of such a misdeed I say: "you could have done otherwise."[16]

First, those words convey a recognition that despite not having acted like someone with the capacity to respond to salient reasons of desirability—to the desiderata that make telling the truth desirable—you did indeed have that capacity at the time of choice: you possessed it, albeit you did not manifest it. Second, the words convey an exhortation after the event to have done otherwise; they communicate that I maintain an attitude that might have been expressed before the event in a regular exhortation to act as the reasons of desirability require. And third, the words convey censure or blame for not having done otherwise; they constitute a reprimand or penalty in themselves—this may be associated, of course, with independent penalties of custom or law—and they communicate at the same time that that penalty is deserved: you do not have an excuse, so I suggest, that might let you off the hook.

According to the first of these connotations, if I say "You could have done otherwise" in response to a misdeed, then I credit you with a capacity to have done otherwise in the situation where you made your choice. This connotation has two elements to it, one negative, the other positive.

The negative connotation is that you were not hindered in either of two commonly recognized ways. First, you were not subject to an agency-debilitating condition like paranoia or obsession or delusion or something of that kind; this radical form of practical excuse would exempt you from being held responsible.[17] And, second, there was no unforeclosed excuse, epistemic or practical, that got in the way of your action. You were in a position to realize that what you said was a lie and that telling a lie in that situation was undesirable. And you were able to act voluntarily on that perception: no one had a gun to your head, for example, and you were not under any other pressure of that kind.

The absence of exemption means that your capacity to respond to the desirability of truth-telling was unimpaired, as it might be said, the absence of excuse that the capacity was unimpeded: there was nothing recognizable in place to block either your recognition of the relevant reasons of desirability or your acting on those reasons. The lack of impairment means that you had the generic capacity to respond to the desirability of truth-telling—to register and act on it—and the lack of impediment means that nothing stopped you from exercising that capacity: you had the situation-specific capacity to respond appropriately. This covers the negative element in the first connotation of holding you responsible. But what does it mean in positive terms to hold that, despite your failure, you had the situation-specific capacity to register the desirability of telling the truth and act as it required?

If you were sensitive to the relevant considerations or reasons of desirability, and you were not affected by the impairment or impediment that might suspend that sensitivity, then your failure to respond appropriately must count as a surprise.[18] Presumably you would have responded appropriately in most variations on that situation where the sensitivity was still in place and there was no impairment or impediment to its activation: that is, to your registering and acting on the considerations. That is the positive element in the first connotation of holding you responsible. It must have been the case, so the presumption goes, that despite the fact that you did not actually respond appropriately, you would have done so over the bulk of variations on the actual situation where the same considerations or reasons continued to obtain and you remained unaffected by exempting or excusing factors.[19] It must have been the case, in short, that it was something of a fluke that you did not register those considerations or act as they required.[20]

There are cases where this connotation of holding you responsible would seem unlikely to be satisfied. Suppose I recognize that you are a habitual liar and that I am not surprised by a lie you just told me. And imagine that still I hold you responsible for the lie, not countenancing any exempting impairment or excusing impediment. How can I seriously believe, in holding you responsible in that way, that just as you were at the time of action, you would have told the truth across most relevant variations on the circumstances? How could I have believed it prior to action in making it clear to you that I would be holding you responsible for acting as the relevant reasons of desirability require? One of the benefits of the narrative presented below is that it makes it intelligible why in Erewhon I might adopt such a view, interacting with you on the assumption that that you are possessed of the appropriate level of capacity. To anticipate, the narrative suggests that not to do so would be to refuse to deal with you within the participant stance of conversation—a stance natural in a society of mutual reliance—preferring instead to treat you in a detached, objective manner as a subject for manipulative treatment.[21]

The second connotation of holding you responsible for a misdeed like telling a lie is that my saying that you could have done otherwise is exhortatory in character. By making this remark, I do not just communicate, as I might communicate to an observer, that as a matter of fact you would generally have done otherwise over variations on the situation where the same reasons were in place and no exemptions or excuses were introduced. And I do not intend to convey just the message that it was a mere fluke that you did not display the disposition in which the capacity consists.[22] Rather I communicate a form of impatience with your failure, a refusal to accept it as a brute fact.

This effect of saying you could have done otherwise means that the remark constitutes a retrospective exhortation, as it might be phrased, to have done otherwise. Normal exhortation is prospective, bearing on a choice that lies before you. I might have exhorted you prior to your choice by saying, "You can respond to the reasons of desirability and tell the truth; you can register and act on those reasons!", where this is meant to support the injunction: "Just do it!" When I say in holding you responsible that you could have done otherwise, I communicate that it would have been appropriate for anyone aware of your situation to have issued such a prospective exhortation prior to the choice. After all, I will communicate that this earlier exhortation would not have been appropriate,

if I concede later that you could not have done otherwise than you did. It is plausible, then, that in saying that you could have done otherwise, I stand by that prior exhortation, whether or not anyone put it to you at the time of choice. In that sense I am naturally taken to exhort you retrospectively to have done otherwise and, by implication, to exhort you to do better in situations of the kind that lie in the future.

The third effect of saying that you could have done otherwise in the case of a misdeed is to censure or reprimand you. Not only do I recognize your capacity to have responded to reasons and told the truth, and not only do I maintain the attitude that I might have expressed earlier by exhorting or enjoining you to tell the truth. I also indict you for the failure to have told the truth. In remarking that you could have done otherwise, I highlight your failure in a presumptively unwelcome manner and thereby reprimand and penalize you for not having told the truth. Moreover I present this reprimand as one that you manifestly deserve; not being able to excuse what you did, it is a reprimand you cannot complain about having to endure.

With these aspects of the responsibility practice spelled out, the question to be explored is whether I and you and other members of Erewhon are likely to hold one another responsible for living up to certain standards of desirability. There is good reason to think that we would evolve this sort of practice. In particular, there is good reason to think that we would come to use the remark, "You could have done otherwise," or some cognate utterance, with the three connotations or effects described.

REGULATING FOR DESIRABILITY

Before developing the argument for this conclusion, however, it is worth noticing that whether or not we evolved the practice of holding one another responsible in Erewhon, we would certainly be likely to regulate one another into responding appropriately to those judgments of unreserved desirability on which we manifestly agreed. We would regulate one another into conformity with such a pattern in the way in which, by the account in the first lecture, we would regulate one another into conformity with a pattern like truth-telling.

In the scenario explored in the first lecture, we have an interest in proving ourselves to be reliable and reputable truth-tellers; unless we do so we cannot expect to be able to rely on others or to get others to rely on us. This interest leads us each to tell the truth in general, seeking to win a

reputation for having the disposition to tell the truth reliably. And that means that just by being there as an audience for one another, ready to make a judgment on whether someone is a careful and truthful speaker, we provide an incentive for one another to tell the truth. We regulate or police one another into conformity with the standard of telling the truth and may be expected to elicit a general pattern of truth-telling.

Suppose now, in line with the evolving narrative, that we share certain standards of unreserved desirability, being responsive to similarly weighted desiderata; that we think there are determinate answers to questions of desirability where those desiderata are the determinants of desirability; and that we each think that anyone who is seeing clearly—anyone not subject to exemption or unforeclosed excuse—will agree with us in the judgment we make in those cases. On the story told so far, we must each reliably respond to such considerations or reasons of desirability—we must recognize and act on their requirements—in the absence of exemption or unforeclosed excuse. If we fail to do so then we cannot expect to be able to rely on others or to get them to rely on us. And that implies that we will each have a reputational incentive to respond appropriately to such reasons.

Absent exemption or unforeclosed excuse, then, we in Erewhon must be expected to regulate or police one another into generally registering the requirements of accepted considerations of desirability and into generally acting as they require. This form of mutual regulation will fall well short of holding one another responsible to those considerations, however, since for all it requires the exercise may not be conscious or intentional. We may regulate one another into responding to the requirements of accepted considerations without any awareness of doing so and without any intention to achieve such an effect. The regulation practiced may be just an unforeseen, aggregate consequence of our individually seeking a reputation for being reliable in our responses.

If I hold you responsible for responding in this sense to accepted considerations of desirability, I do something much richer than anything I need do in policing you in this way. Thus if I blame you for not acting as the considerations required, I will normally blame you consciously and intentionally. This will certainly be so if I express the blame in words, as in saying, "You could have done otherwise." But it will also be the case if I assume an attitude of blame and keep it to myself. It barely makes sense—although it may convey something metaphorically—to imagine that I might blame you but only unconsciously or unintentionally.

But while regulating one another for responding to accepted reasons of desirability falls short of holding one another responsible to such reasons, the regulative regime may still play an important role in supporting the responsibility practice. This will appear in the story to be told of how we in Erewhon might come to hold one another responsible. The narrative assumes that we are subject to a reputational discipline in which, as a matter of common awareness, we expect one another to be responsive to the expectations of reliability—in particular, reliability in responding to the requirements of accepted considerations of desirability—that we elicit or license in one another.

It should be no surprise that the practice of holding one another responsible depends on the presence of a basic regulative infrastructure of this kind, for that practice itself has a clear regulative rationale. In recognizing your capacity to have acted on relevant reasons, in exhorting you retrospectively to have done so, and in censuring you for your failure, it should be clear that I am working with the assumption that I can thereby influence and even reform you. I may not blame you with an explicitly reformative intention: my primary intention may be just to draw attention to your failure to respond to what by shared lights are the demands of desirability. But there would scarcely be any point in holding you responsible for such failures, if I thought that there was no possibility of getting you to change.[23]

It is time now to return to the narrative and explain why we in Erewhon might go beyond blind regulation and hold one another responsible for living up to accepted reasons of desirability; in particular, to show how we might begin to use something like the remark, "You could have done otherwise," with the three effects associated with holding you responsible. It will be enough to show that such a remark would naturally have those effects, uttered within the Erewhonian world where the concept of unreserved desirability has gained a hold. For if it can saliently have such effects, and thereby implement a system of mutual regulation, then that will give each of us a motive for making the utterance in response to this or that misdeed.

The effects to be explored are the effect of recognizing the offender's capacity to have done otherwise; the effect of exhorting the offender retrospectively to have done otherwise; and the effect of reprimanding the offender for the failure, communicating the message that the reprimand is deserved: there is nothing they can say to excuse themselves. These may

be described respectively as the recognition effect, the exhortation effect, and the censure effect.

The Recognition Effect

There are two possible readings of the remark "You could have done otherwise" that I might utter in Erewhon, responding to an offence against some standards of desirability: presumptively, standards that I take you to agree with me in endorsing. On one reading it would mean, in the sense explicated, that you had the capacity to do otherwise: sticking with our example, that you had the capacity to respond to accepted reasons of desirability and tell the truth. On this reading, strictly taken, it would mean that you were disposed in the situation of choice to respond robustly to those reasons—to respond to them in any situation similar to the actual circumstances in which the reasons were present and there was no exemption or unforeclosed excuse—and that your failure to do so was a fluke. On a rival reading of the remark, however, it would mean just that you would have done otherwise if you were a different sort of person: that you would have told the truth if you had had a better education, for example, or had not lived in bad company for so long.

Why would the remark attract the first reading, and have the default effect of communicating the recognition of a capacity to respond appropriately to relevant reasons? Why would I not be moved by the evidence of your failure to conclude that actually you were not responsive to the requirements of shared standards of desirability: you were not disposed, just as you were, to register the relevance of those considerations and to act as they required?

It would certainly be reasonable to ignore the evidence of a particular failure if you had already demonstrated that capacity over a range of similar cases. But it is a default assumption in the practice of holding one another responsible that, even in the absence of such a record, the person who offends against accepted standards of desirability—assuming there is no exemption and no unforeclosed excuse—is fit to be held responsible for the offence and so must have had a capacity to respond to relevant reasons in the exercise of the choice. Is it possible to explain why we who live in Erewhon might be led to support a default assumption of this kind? Plausibly, it is.

Erewhon is a world where we each expect in our own case that others will rely on our words, when we expose ourselves to their scrutiny and

their sanction and are not blocked from living up to those words by any exemption or any unforeclosed excuse. We expect that others will act on the assumption that our subjection to that reputational discipline will help to ensure our reliability. This means that it must be a matter of common assumption in Erewhon that the reputational discipline we invite and impose on one another is sufficient, in the absence of exemption or unforeclosed excuse, to ensure an important result: the presence of capacities to conform to patterns we routinely endorse. The observation holds, as registered earlier, not just with conforming to a pattern like truth-telling, but also with conforming to a pattern like that of responding reliably to accepted reasons of desirability.

Suppose now that I take you to accept the reasons of desirability that require you to tell the truth but that you actually tell a lie and do so in the absence of exemption or unforeclosed excuse. To judge that you did not have the capacity to respond to those reasons in the situation where you acted would be to treat you as someone beyond the reputational discipline within which we relate to one another. It would be to suspend the assumption of capacity that operating under that discipline supports and to give up on you in resignation or despair. Thus, if I take you as a fellow subject of that discipline, I must be disposed to treat you as possessed of that capacity and to think that you committed the offence simply because you failed to exercise the capacity.

This argument is worth spelling out more carefully. Taking you to be subject to the reputational discipline that characterizes our relationships in Erewhon, I am bound to think that your exposure to the expectations of others—in particular, expectations that you elicit or license—is likely to provide you with a powerful motive to live up to them and, in that sense, to establish a capacity to do so. But insofar as you manifestly endorse considerations of desirability that clearly argued for telling the truth in the situation in which you acted—insofar as you co-avow or co-accept those considerations—it is clear that I and others would have expected you to live up to them by telling the truth. And that means that it is equally clear, and certainly clear by my lights, that you had the motive and capacity to do so. Thus I must think that you failed to live up to those considerations and tell the truth in the presence of a capacity to have done so, and not that the failure was due to the absence of such a capacity.

This is to say that in the case considered I have epistemic grounds, albeit grounds of an unusual sort, to ascribe a capacity to have done otherwise to you. The grounds are not that you had the capacity, just as you

were in yourself, to respond appropriately to the relevant considerations. Rather they are that you had the capacity, as someone empowered by the reputational culture in which we are commonly immersed, to have responded appropriately and told the truth.

These grounds will be available to me, of course, only insofar as I refuse to treat you as an outside or outlier: someone beyond the reputational community. But it makes good practical sense to assume by default that you are not someone of this kind. In Erewhon, as characterized in the narrative, the very possibility of cooperation and community depends on our each making a default assumption in dealing with one another that our words are our bonds and, more generally, that we can be relied upon to live up to the expectations that we elicit or license in others.[24] To reject that default in dealing with you would be to deny in effect that you were one of us.

I might be driven to withdraw the assumption in your particular case, of course—I might be forced to treat you as a pathological liar, for example—if your failure was repeated time and time again. But the cost of ostracizing you in this way would be enormous, since it would mean giving up on the possibility of our having a reputational influence on you within the community. It would make little sense to adopt such an attitude of hopeless resignation in light of a single offence, or even a limited record of offence. Doing so would be a resort of utter despair.

The Exhortation Effect

If these considerations are sound, then when I say that you could have done otherwise in wake of a misdeed, in particular some misdeed where you were not subject to an exempting or an unforeclosed excusing condition, then I should be taken to convey by those words that you had the capacity to do otherwise. You were someone disposed reliably, if not invariably, to respond to reasons of desirability and act as they require. On the account offered earlier, however, those words should convey a second message too, if they are to represent an instance of holding you responsible. They should communicate that I think it would have been appropriate for anyone to exhort and enjoin you, prior to the action, to respond appropriately to those reasons. In that sense they should have the effect of a retrospective exhortation.

Suppose, then, that I say that you could have done otherwise in wake of a misdeed such as telling a lie, where it is granted that the action offends against recognized standards of desirability, and that it was performed in

the absence of an exemption or an unforeclosed excuse. Is there any reason to think that in Erewhon these words would naturally have a retrospectively exhortatory significance? For reasons related to the reputational discipline just invoked, it turns out that there is.

In Erewhon, as already argued, we each make good use of the reputational pressures that others bring to bear on us when they form expectations, by our license, about what we will think and do. Those pressures force us to be careful to advertise only attitudes we can live up to and to be careful about living up to the attitudes we advertise. They ensure, in effect, that we have resources enough to establish ourselves as reliable partners and neighbors, giving us each capacities that we might not have had in their absence. In particular, they give us the capacity to respond to accepted reasons of desirability, enabling us in the absence of exemption or unforeclosed excuse to register what they require in any instance—say, to tell the truth—and to act on that requirement.

On this account, the capacity any one of us has to respond to reasons of desirability is liable to depend not just on our own internal powers, but on the social or reputational environment in which we operate; it is likely to have an ecological character.[25] This lesson is evidentially available to all of us, as is that availability itself, the availability of that availability, and so on. And so we are likely to hold it as a matter of common awareness in the community.

But if this is a matter of common awareness, then it must be equally a matter of common awareness that when I speak to you prior to your making a choice and say that there are reasons of desirability to do something such as telling the truth, or just that you can tell the truth, then I assume a role in which I may expect to have an influence on you. I am in a position to speak, not just in the manner of someone recording your capacity to tell the truth—say out of a concern for historical accuracy—but also in the manner of someone consciously hoping to elicit that capacity in the very act of ascribing it.

I do not speak just descriptively in saying or implying that you have the capacity, then, as I might do in saying that you have a ruddy complexion; I do not record a situation that obtains independently of what I say. Nor of course do I speak performatively, as I might do in saying "I resign"; I do not record a situation that is made to be true by the very words I utter.[26] I speak evocatively, so it might be put, using words that serve at once to record a situation—your having the capacity to tell the truth—and to make it more likely to obtain. I speak with the manifest

expectation, and presumably the intention, of evoking the very capacity I ascribe.[27]

Suppose, then, that when I say "You can tell the truth," it is generally understood that I am likely to be speaking in this evocative manner, exhorting you to tell the truth and supporting by implication the injunction to tell the truth. What does that imply for how I am likely to be speaking when I say in the wake of your failure that you could have done otherwise: you could have told the truth? Plausibly, it implies that I am almost certainly speaking in the manner of retrospective exhortation.

Let "You can tell the truth" have the force of an exhortation when uttered prior to a choice. That is more or less bound to ensure that "You could have told the truth," uttered in the wake of a choice, is going to communicate the message that despite your failure, it would have been appropriate to exhort and enjoin you to tell the truth prior to the choice. And that will be so whether or not I or anyone else actually issued the prior exhortation. The remark will communicate that I maintain the attitude that might have been expressed earlier by "You can tell the truth." And so, as the second effect requires, it represents a form of retrospective exhortation.

The Censure Effect

The observations so far show that having developed along the lines charted, I and you and others in Erewhon would satisfy the first two conditions associated with holding someone responsible. Thus I would be in a position, absent considerations of exemption or unforeclosed excuse, to give default recognition to your capacity to respond to the reasons of desirability that required to tell the truth, even in the wake of failure. And I would be in a position to speak with an exhortatory, injunctive force in saying in the wake of any such failure that you could have done otherwise. The final question is whether I could also be taken to censure you by making such a remark, imposing the penalty of a reprimand and implying at the same time that this penalty—and perhaps an associated form of punishment—is deserved.

By the account developed so far, my saying you could have done otherwise in the wake of a misdeed—say, a lie—presupposes that it was manifestly appropriate for anyone prior to your action to exhort and enjoin you to respond to reasons of desirability and tell the truth; by assumption, the option of telling the truth was unreservedly desirable in the circumstances, by our shared lights. But if it was manifestly appropriate for

anyone to have enjoined you to respond to reasons of desirability and to tell the truth, then in the wake of the failure, it is manifestly appropriate for me to register that you acted in violation of such an injunction. And that is something I can be plausibly taken to do in saying that you could have done otherwise. In the context, this amounts to registering that you did not act as it would have been appropriate for anyone to enjoin you to act. In saying that you could have done otherwise, I stand by the appropriateness of the injunction and mark your failure to satisfy it.

This in itself is to impose a recognized penalty on you. For it is to express a bad opinion of your failure to act as you might appropriately have been enjoined to act. In effect, it is to issue a reprimand for the way you behaved. And not only does the remark constitute a reprimand; it also communicates that the reprimand is itself deserved. In saying that you could have done otherwise, conveying the message that you acted against an appropriate injunction, I indicate that you were not subject to an exempting or unforeclosed excusing condition; its presence would have meant that in a relevant sense you could not have done otherwise than you did. And in indicating the absence of such factors, I emphasize that there is nothing, under our practices, that might lead me to withdraw the reprimand. I put you on the hook and, since you do not have any available excuse for what you did, you cannot complain about my doing so.

CENSURE IN A NATURALISTIC WORLD

Anti-naturalists hold that in order for you to deserve a reprimand—in order for you to count as blameworthy—it must be the case, not only that there was no available excusing or exempting factor at the origin of your action, but that there was no regular causal factor whatsoever at its source. The idea is that in order to be blameworthy the action must have issued from an uncaused will. It must be something that you brought about as an agent, not something that was occasioned within you, say by a chance failure of normal functioning. The action must have been up to you, and only up to you, in a sense that rules out naturalistic causation.

On standard naturalistic premises, such as those assumed in these lectures, there are no events that cannot be traced to natural causal or chance antecedents. And so it is particularly important on this approach that the practice of holding responsible should not imply that if some causal or chance factor affected your performance, you are off the hook. Happily, however, the practice that emerges in Erewhon can make perfect sense, even if naturalism is sound. The narrative shows that you can

be blameworthy in a significant sense even if the action for which I censure you can be causally traced to some dysfunctional blip or glitch, or just to brute chance.

The exhortatory, injunctive practice that obtains in Erewhon appears and survives, by the narrative adopted, because of a wish on the part of members to increase their perceived reliability by exposing themselves to costs in the case of failure: in particular, a failure to live up to the requirements of accepted standards of desirability. The practice allows that if a special set of causes—those associated with exemptions and unforeclosed excuses—can be adduced to explain a failure, then you do not incur those costs and cannot deserve a reprimand; you are off the hook. But assuming that there are causes of failure apart from these—or assuming that chance can play a role in generating failure—the practice cannot allow such factors to let you off the hook. Otherwise it would lose its regulatory point and frustrate the wish that supports it. No practice of holding people responsible to certain standards could have an impact on their performance if offenders could get off the hook just by arguing that their offence was the effect of a natural cause.

Assume that in Erewhon we treat exempting and unforeclosed excusing factors as having the following feature: that if they are present, then even the costs that we face for failing to live up to our advertised attitudes are not going to be enough to get us to display those attitudes. And assume in addition that we think that the other factors that might occasion such a failure—factors like the neural blip or glitch, of which we may know little—are different in precisely that respect: even if they are present, the costs that we face for not living up to our advertised attitudes are sufficient to trump them, although perhaps only in the light of experience and education. In a phrase, assume that in suitable contexts we treat exempting and unforeclosed, excusing factors as ones we cannot regulate one another into overcoming and that we treat other causal factors as ones we can; we treat the former as resistant to the effects of regulation, the latter as susceptible to those effects.

Under these assumptions, it will make perfect sense in Erewhon for us to treat offences that derive from regulation-resistant factors as not deserving blame and offences that derive from regulation-susceptible factors as deserving blame. And of course it will make sense for us to be open to experience in determining which items should be put in the resistant category, and which in the susceptible. It may be true, as noticed earlier, that the practice of holding one another responsible is distinct

from the practice of blind regulation. But in the approach suggested there is a deep and continuing connection between the two. It is ultimately because of its regulative point that the practice of holding one another responsible can make the distinction between misdeeds that are deserving of blame and those that are not.[28]

Why, then, should you pay the costs associated with a misdeed that is due to a regulation-susceptible factor: say, the unknown neural blip or glitch? Why should you be expected to treat a reprimand as deserved? In a word, because the glitch counts as an influence that you are able, by the common sense of Erewhonians, to overcome: you have all the motivation and resolution required to carry you past it, especially given the force that you unleash in exposing yourself to the possibility of reputational loss. Factors that count as regulation-resistant ones that you are unable to overcome in the same way. They are not special because they obstruct the operation of an allegedly uncaused will; any causes would serve to do that. They are special because they stand out among natural causes by virtue of the fact that there is not much that you can do, no matter how motivated you are, to overcome them.[29]

This observation take us to the denouement. Just as the developments charted in the present narrative make sense of why I and you and others in Erewhon should give one another the recognition and exhortation associated with the practice of holding responsible, so they make sense of why we should also impose the censure associated with that practice. Thus, the narrative not only explains why we would develop the concept of desirability, it also makes sense of why we would begin to hold one another responsible for living up to certain standards of unreserved desirability. It shows that we would be led, as by an invisible hand—and not, for example, as the result of planning or contract—to make certain judgments of desirability and to hold one another responsible for acting according to those judgments.

BACK TO RIGHTNESS

Before concluding the discussion, however, there is one loose end to tie up. By most accounts it is the concept of the right or obligatory that is central to ethics, not the concept of the desirable. So how does it fit into our picture?

On the line adopted earlier, it is right or obligatory for an agent to choose one option rather than another if and only if, first, that option is unreservedly desirable and, second, the agent would be blameworthy for

not choosing it. Thus if we Erewhonians are disposed in many cases to hold one another responsible for failures to do what is unreservedly desirable, then the concept of the right or obligatory will be within our ready reach. We will be prepared to hold that it would have been right for an agent to select a certain option just in case that option was unreservedly desirable and the agent can be held responsible for a failure to perform it, attracting the sort of recognition, exhortation, and censure just discussed.

This is not to say that the category of the right or obligatory would take over completely from that of the desirable. Certain options might be unreservedly desirable, by standards accepted across the community, without counting as right or obligatory. They might require such a level of sacrifice that we would not be prepared to blame people for failing to perform them: we would balk in the case of an offence at ascribing to them the capacity to have responded to the relevant standards. Hence there would be room for a divergence between the category of the unreservedly desirable and the right or obligatory. Those options that are not obligatory would count, in the received term, as supererogatory.

Conclusion

It may be useful in conclusion to remind ourselves of the main steps taken in this second lecture.

- The aim in explaining the emergence of ethics in Erewhon is to account for how we inhabitants could come to think in terms of desirability and responsibility.
- On the desirability front it is essential to explain how we could come to think of one option in a set of alternatives as unreservedly desirable, not just desirable under a certain aspect.
- From within the perspective of avowal, we Erewhonians are bound to distinguish between what we actually believe or desire and what our avowals require us to believe or desire.
- Our avowals require us to form beliefs robustly on the basis of data and to form desires robustly on the basis of desiderata, resisting any pressures that count as potential distorters by practice-related criteria.
- From within the practice of avowal, then, we must treat robustly persuasive propositions as credible or ought-to-be-believed, robustly attractive scenarios as desirable or ought-to-be-desired, counting it as a failure if our actual beliefs and desires diverge.

- The practice of co-avowal may involve co-avowing in the name of an open or a closed group; it gives rise to a concept of the commonly credible or desirable in the first case, the jointly credible or desirable in the second.
- In the case of credibility, the commonly credible naturally figures as a master category: what any one of us ought to believe in light of the practice of open co-avowal, we ought to believe unreservedly.
- In the case of desirability, no category has this master status and it is essential for us to have a basis for resolving issues about what to choose in cases where different options are individually, commonly, and jointly desirable.
- The natural way for us to do this is to allow for the aggregation of desiderata across the practices associated with these forms of desirability and the identification in a problematic choice of an option that is practice-neutrally desirable.
- The concept of the practice-neutrally desirable satisfies the constraints governing the ordinary concept of the unreservedly desirable and ensures the success of the genealogy on this front.
- But would we Erewhonians hold one another responsible for living up to the requirements of accepted standards of unreserved desirability? That is the second challenge for a genealogy of ethics.
- Holding you responsible for a misdeed involves ascribing a capacity not to have done it, exhorting you retrospectively to have done it, and reprimanding and penalizing you for failing.
- In Erewhon we rely on one another to impose a reputational discipline that enables us each to live up to others' expectations and it is a matter of common belief that we enjoy reputationally enhanced capacities on this front.
- Thus it is going to be a default assumption among us that if you offend against accepted standards of desirability in the absence of exemption or unforeclosed excuse, I can reasonably credit you with the capacity to have done otherwise.
- For similar reasons I can hold that you might have been appropriately exhorted before acting to take the unreservedly desirable option and I can reaffirm that exhortation retrospectively by saying "You could have done otherwise."
- Finally, by uttering those words I can mark your failure in an unwelcome way, thereby reprimanding and penalizing you; and I can do this in conditions where it is manifest that you cannot object to my doing so.
- With this account of how we Erewhonians might have come to think in terms of desirability and responsibility, it is also possible to make room for the right or obligatory. An option will be right just in case it is unreservedly desirable and the failure to enact it would be blameworthy.[30]

NOTES

1. This may also be described as what you ought to desire, all things considered or, in John Broome's phrase, *pro toto* (John Broome, *Rationality through Reasoning* [Oxford: Wiley Blackwell, 2013]). This phrasing not only marks the contrast with what you ought to desire *pro tanto* but also points to the explanation for why you ought to desire something *simpliciter*: namely, because of its properties overall.
2. Huck Finn does well overall by not letting his slave-culture sense of what is unreservedly desirable affect his spontaneous desire to help out the enslaved Jim. But he still displays a form of failure in not being guided by the judgment of desirability to which that sense of things leads him. It is just that that failure is a happy fault: it compensates for the worse failure that consists in his sense of unreserved desirability being warped by the slave culture and his judgment of desirability being therefore mistaken. That sense of desirability gives him a poor map and it is just as well that he does not navigate by it: that he relies instead, as we might say, on an intuitive analogue of dead reckoning.
3. In this second case, "unreservedly desirable" works the way "prudent" does. As it is prudent for you to look after your future, for me to look after mine, it may be unreservedly desirable for you to favor your child, unreservedly desirable for me to favor mine. It is important to register that "unreservedly desirable," like "prudent," may vary between contexts in the property it ascribes but that nevertheless it is not indexical in character. The point also applies to "right," used in a corresponding way. I am grateful to John MacFarlane for discussion of this issue.
4. John MacFarlane, *Assessment Sensitivity: Relative Truth and Its Applications* (Oxford: Oxford University Press, 2014).
5. R. M. Hare, *The Language of Morals* (Oxford: Oxford University Press, 1952).
6. Frank Jackson and Philip Pettit, "Moral Functionalism and Moral Motivation," *Philosophical Quarterly* 45 (1995): 20–40; reprinted in Frank Jackson, Philip Pettit, and Michael Smith, *Mind, Morality and Explanation* 189–210 (Oxford: Oxford University Press, 2004).
7. Lewis, *Convention*.
8. Consider the measures that are strategically attractive for any one of us in Erewhon, as tracked in the first lecture: measures like telling the truth and establishing ourselves as reliable reporters. We might have seen these as individually desirable, if we had had access to that concept. But there is no reason to think that we must have access to that concept in order to behave strategically. As noted earlier, the genealogy presented does not assume that we Erewhonians had access to the concept of the individually desirable–or to any desirability concepts–at the purely reportive stage charted in the first lecture.
9. The practice requires these responses in the way in which we say that the law requires us to behave thus and so. On this sense of how a practice can be a source of requirements, see Broome, *Rathionality through Reasoning*, ch. 7.
10. This is to say that the ideal self may advise that the actual self should behave in a manner that takes account of difficulties the ideal self does not itself have to

deal with. On this lesson, which also applies in the idealizations considered later, see Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994).

11. It is perhaps something like this that Richard Rorty has in mind when he appeals to the conversation of humanity. Richard Rorty, *Philosophy and the Mirror of Nature* (Oxford: Basil Blackwell, 1980).
12. This is not to deny that there are some standpoints, say those associated with a form of oppression to which I am personally not subject, such that it may take enormous efforts of empathy on my part—and a willingness to trust the testimony of those occupying that standpoint—for me to grasp what is revealed therein. Similar points apply, of course, in the case of the commonly desirable. See Karen Jones, "Second-hand Moral Knowledge," *Journal of Philosophy, 96* (1999): 55–78.; and for a general perspective, see Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press, 2007).
13. Christian List and Philip Pettit, *Group Agency: The Possibility, Design and Status of Corporate Agents* (Oxford: Oxford University Press, 2011); and Lara Buchak and Philip Pettit, "Reasons and Rationality: The Case for Group Agents," in *Weighing and Reasoning*, edited by I. Hirose and A. Resner, 207–31 (Oxford: Oxford University Press, 2014).
14. This observation may offer some support for the pragmatist thought that the true is that which is destined to be agreed upon.
15. Their adversaries in such a case will take it to be commonly credible—and so credible, period—that the relevant actions are unreservedly desirable. And so they will take the amoralists to reject the claims of the commonly credible without offering any persuasive rejoinder.
16. To hold you responsible for a misdeed is not necessarily to make any utterance. But it is to assume the attitude that might be expressed by saying that you could have done otherwise or had reason to do otherwise or something of the kind.
17. R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge, MA: Harvard University Press, 1996); and John Gardner, *Offences and Defences: Selected Essays in the Philosophy of Criminal Law* (Oxford: Oxford University Press, 2007).
18. This will be a surprise, of course, only if I am truly holding you responsible, not pretending to hold you responsible. Thus it will not be a surprise if I am going through the motions of holding you responsible, because of endorsing the assumption, not that you have the capacity to do otherwise, but that you are capable of developing that capacity as a result of being treated as if you already had it. This would not be to treat you as responsible, strictly speaking, but to treat you as "responsibilizable": capable of being made fit to be held responsible by being held responsible, when strictly you lack such fitness. On that theme, see Philip Pettit, "Responsibility Incorporated," *Ethics* 117 (2007): 171–201.
19. Michael Smith, "Rational Capacities, or: How to Distinguish Recklessness, Weakness and Compulsion," in *Weakness of Will and Practical Irrationality*, edited by S. Stroud and C. Tappolet, 17–38 (Oxford: Oxford University

Press, 2003); and Victoria McGeer and Philip Pettit, "The Hard Problem of Responsibility," *Oxford Studies in Agency and Responsibility*, vol. 3, edited by D. Shoemaker, 160–88 (Oxford: Oxford University Press, 2015).

20. There are two dimensions to the notion of a degree of capacity, which might be described as dependability and durability. A disposition to respond to reasons of desirability will score higher in dependability the more likely it is to survive temptations, and score higher in durability the more likely it is to survive disrupters. Where temptations have to be registered as considerations that counter reasons of desirability, disrupters need not be: like distractions or mood swings they can operate behind the back of the agent. See Pettit, *The Robust Demands of the Good,* ch. 2, and Victoria McGeer and Philip Pettit, "The Empowering Theory of Trust," in *The Philosophy of Trust,* edited by Paul Faulkner and Thomas Simpson (Oxford: Oxford University Press, 2016). Only dependability is taken into account in the text.
21. See Peter Strawson, *Freedom and Resentment and Other Essays* (London: Methuen, 1962) and Philip Pettit and Michael Smith, "Freedom in Belief and Desire." *Journal of Philosophy* 93 (1996): 429–49; reprinted in *Mind, Morality and Explanation*, edited by Frank Jackson, Philip Pettit, and Michael Smith, 375–96 (Oxford: Oxford University Press, 2004).
22. Pamela Hieronymi, "Rational Capacity as a Condition on Blame," *Philosophical Books* 48 (2007): 109–23.
23. Strawson, *Freedom and Resentment and Other Essays;* Victoria McGeer, "Strawson's Consequentialism," *Oxford Studies in Agency and Responsibility* 2 (2014): 64–92.
24. We may elicit or license suitable expectations by what we positively say but also by what we fail to say: for example, by our failure to reject the manifest assumption on the part of others that we endorse certain judgments of unreserved desirability.
25. Victoria McGeer, "Civilizing Blame," in *Blame: Its Nature and Norms*, edited by J. D. Coates and N. A. Tognazzini, 162–88 (Oxford: Oxford University Press, 2013); and McGeer and Pettit, "Hard Problem of Responsibility." The idea of an ecological capacity is borrowed from Manuel Vargas, *Building Better Beings: A Theory of Moral Responsibility* (Oxford: Oxford University Press, 2013). Using concepts introduced earlier, a capacity might depend on ecology for being dependable or for being durable or both.
26. David Lewis, *Philosophical Papers*, vol. 1 (Oxford: Oxford University Press, 1983), ch. 12.
27. McGeer and Pettit, "Hard Problem of Responsibility."
28. I am particularly indebted to Victoria McGeer for the line in this paragraph, as I am indebted on the topic of responsibility in general to our joint work (ibid.).
29. The idea that you might be willing to accept the penalty for a failure that has a natural cause, provided that cause is not a recognized excuse, is borne out by the way we treat causes of moral failure such as laziness or weakness of will. Neither of these factors counts in ordinary terms as an excuse. And yet each is frequently invoked as a cause of the failure to act appropriately. Why do we

treat laziness and akrasia as causes of failure but not causes of the sort that might excuse it? Plausibly, it is because we think of them as hurdles that people of normal motives and resources are capable of overcoming.

30. I presented lectures based on this text at the Australian National University in March 2015 and then at the Tanner Lectures in Berkeley in early April 2015. I learned enormously from the many comments received at those events and afterward and am in the debt of too many people to list. However, I must mention my Tanner commentators, Pamela Hieronymi, Richard Moran, and Michael Tomasello. Apart from offering enlightening commentaries on the material, to which I hope to respond in a separate publication, they were most generous with their informal advice and criticism.

REFERENCES

Alonso, F. M. "What Is Reliance?" *Canadian Journal of Philosophy* 44 (2014): 163–83.

Anscombe, G. E. M. *Intention*. Oxford: Blackwell, 1957.

Appiah, Kwame Anthony. *The Honor Code: How Moral Revolutions Happen*. New York: Norton, 2010.

Axelrod, Robert. *The Evolution of Cooperation*. New York: Basic Books, 1984.

Ayer, A. J. *Language, Truth and Logic*. London: Gollanz, 1982.

Bar-on, Dorit. *Speaking My Mind: Expression and Self-knowledge*. Oxford: Oxford University Press, 2004.

Bennett, Jonathan. *Rationality*. London: Routledge and Kegan Paul, 1964.

Blackburn, Smon. *Spreading the Word*. Oxford: Oxford University Press, 1984.

Boehm, Christopher. *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press, 1999.

Boix, Carles, and Frances Rosenbluth. "Bones of Contention: The Political Economy of Height Inequality." *Americal Political Science Review* 108 (2014): 1–22.

Bratman, Michael. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press, 2014.

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood. *Explaining Norms*. Oxford: Oxford University Press, 2013.

Brennan, Geoffrey, and Philip Pettit. *The Economy of Esteem: An Essay on Civil and Political Society*. Oxford: Oxford University Press, 2004.

Broome, John. *Rationality through Reasoning*. Oxford: Wiley Blackwell, 2013.

Byrne, Alex, "Transparency, Belief, Intention," *Proceedings of the Aristotelian Society* 85 (2011): 201–21.

Buchak, Lara, and Philip Pettit. "Reasons and Rationality: The Case for Group Agents. In *Weighing and Reasoning*. Edited by I. Hirose and A. Resner. Oxford: Oxford University Press, 2014.

Chalmers, David, and Frank Jackson. "Conceptual Analysis and Reductive Explanation." *Philosophical Review* 110 (2001): 315–60.

Coleman, James. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press, 1990.

Craig, Edward. *Knowledge and the State of Nature*. Oxford: Oxford University Press, 1990.
Dancy, Jonathan. *Ethics without Principles*. Oxford: Oxford University Press, 2004.
Davidson, Donald. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 1984.
DeScioli, Peter, and Robert Kurzban. "A Solution to the Mysteries of Morality." *Psychological Bulletin* 139 (2013): 477–96.
Dennett, Daniel. *Brainstorms*. Brighton: Harvester Press, 1979.
Dietrich, Franz, and Christian List. "A Reason-Based Theory of Rational Choice." *Nous* 47 (2013): 104–34.
Elster, Jon. *Alchemies of the Mind: Rationality and the Emotions*. Cambridge: Cambridge University Press, 1999.
Evans, Gareth. *The Varieties of Reference*. Oxford: Oxford University Press, 1982.
Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007.
Gardner, John. *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*. Oxford: Oxford University Press, 2007.
Gert, Joshua. *Normative Bedrock: Response-dependence, Rationality, and Reasons*. Oxford: Oxford University Press, 2012.
Gibbard, Allan. *Wise Choices, Apt Feelings*. Oxford: Oxford University Press, 1990.
Gilbert, Margaret. *Joint Commitment: How we Make the Social World*. Oxford: Oxford University Press, 2015.
Grice, Paul. "Logic and Conversation." In *Syntax and Semantics*, vol 3. Edited by P. Cole and J. L. Morgan. New York: Academic Press, 1975.
———. "Method in Philosophical Psychology." *Proceedings and Addresses of the American Philosophical Association* 68 (1975): 23–53.
———. *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press, 1989.
Hare, R. M. *The Language of Morals*. Oxford: Oxford University Press, 1952.
Hart, H. L. A. *The Concept of Law*. Oxford: Oxford University Press, 1961.
Hieronymi, Pamela. "Rational Capacity as a Condition on Blame." *Philosophical Books* 48 (2007): 109–23.
Hobbes, Thomas. *Leviathan*. Edited by E. Curley. Indianapolis: Hackett, 1994.
Hume, David. *Political Essays*. Cambridge: Cambridge University Press, 1994.
Jackson, Frank. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press, 1998.
Jackson, Frank, and Philip Pettit. "Moral Functionalism and Moral Motivation." *Philosophical Quarterly* 45 (1995): 20–40; reprinted in Frank Jackson, Philip Pettit and Michael Smith. *Mind, Morality and Explanation*. Oxford: Oxford University Press, 2004.
———. "A Question for Expressivists." *Analysis* 58 (1998): 239–51.
Jackson, Frank, Philip Pettit, and Michael Smith. "Ethical Particularism and Patterns." In *Particularism*. B. Hooker and M. Little, 1999.

Jones, Karen. "Second-hand Moral Knowledge." *Journal of Philosophy* 96 (1999): 55–78.
Joyce, Richard. *The Evolution of Morality*. Cambridge, MA: MIT Press, 2006.
Kitcher, Philip. *The Ethical Project*. Cambridge, MA: Harvard University Press, 2011.
Langton, Rae. *Sexual Solipsism: Philosophical Essays in Pornography and Objectification*. Oxford: Oxford University Press, 2009.
Lewis, David. *Convention*. Cambridge, MA: Harvard University Press, 1969.
———. *Philosophical Papers*, vol 1. Oxford: Oxford University Press, 1983.
List, Christian, and Philip Pettit. *Group Agency: The Possibility, Design and Status of Corporate Agents*. Oxford: Oxford University Press, 2011.
Lyon, A. *Resisting Doxastic Pluralism: The Bayesian Challenge Redux*. University of Maryland. College Park, 2015
MacFarlane, John. *Assessment Sensitivity: Relative Truth and its Applications*. Oxford: Oxford University Press, 2014.
Mackie, J. L. *Ethics*. Harmondsworth: Penguin, 1977.
Maynard Smith, John, and David Harper. *Animal Signals*. Oxford: Oxford University Press, 2004.
McGeer, Victoria. "Civilizing Blame." In *Blame: Its Nature and Norms*. Edited by J. D. Coates and N. A. Tognazzini, 162–88. Oxford: Oxford University Press, 2013.
———. "Is 'Self-knowledge' an Empirical Problem? Renegotiating the Space of Philosophical Explanation." *Journal of Philosophy* 93 (1996): 483–515.
———. "The Moral Development of First-Person Authority." *European Journal of Philosophy* 16 (2008): 81–108.
——— "Strawson's Consequentialism." *Oxford Studies in Agency and Responsibility* 2 (2014): 264–92.
McGeer, Victoria, and Philip Pettit. "The Empowering Theory of Trust." In *The Philosophy of Trust*. Edited by Paul Faulkner and Thomas Simpson. Oxford: Oxford University Press, 2016.
———. "The Hard Problem of Responsibility." In *Oxford Studies in Agency and Responsibility*, vol. 3. Edited by D. Shoemaker, 160–88. Oxford: Oxford University Press, 2015.
———. "The Self-regulating Mind." *Language and Communication* 22 (2002): 281–99.
Menger, Carl. "On the Origin of Money." *Economic Journal* 2 (1892): 239–55.
Moran, Richard. *Authority and Estrangement: An Essay on Self-knowledge*. Princeton, NJ: Princeton University Press, 2001.
———. "Self-Knowledge: Discovery, Resolution, and Undoing." *European Journal of Philosophy* 5 (1997): 141–61.
Neuhouser, Frederick. *Rousseau's Critique of Inequality: Reconconstructing the Second Discourse*. Cambridge: Cambridge University Press, 2015.
Nietzsche, Friedrich. *On the Genealogy of Morals*. Cambridge: Cambridge University Press, 1997.
Pettit, Philip. *The Common Mind: An Essay on Psychology, Society and Politics*. New York: Oxford University Press, 1993.

———. "Decision Theory and Folk Psychology." In *Essays in the foundations of Decision Theory*. Edited by M. Bacharach and S. Hurley. Oxford: Blackwell, 1991; reprinted in Philip Pettit. *Rules, Reasons, and Norms*. Oxford: Oxford University Press, 2002.

———. "Group Agents are not Expressive, Pragmatic or Theoretical Fictions." *Erkenntnis* 79 (2014).

———. *Made with Words: Hobbes on Language, Mind and Politics*. Princeton, NJ: Princeton University Press, 2008.

———."Making Up Your Mind." *European Journal of Philosophy* 23 (2015).

———. "Practical Belief and Philosophical Theory." *Australasian Journal of Philosophy* 76 (1998): 15–33.

———. "Responsibility Incorporated." *Ethics* 117 (2007): 171–2001.

———. *The Robust Demands of the Good: Ethics with Attachment, Virtue and Respect*. Oxford: Oxford University Press, 2015.

———. Value-mistaken and Virtue-mistaken Norms. *Political Legitimization without Morality?* Edited by Jörg Kühnelt, 139–56. New York: Springer, 2008.

———. "*Virtus Normativa*: Rational Choice Perspectives." *Ethics* 100 (1990): 725–55; reprinted in Philip Pettit, *Rules, Reasons, and Norms*. Oxford: Oxford University Press, 2002.

Pettit, Philip, and David Schweikard. "Joint Action and Group Agency." *Philosophy of the Social Sciences* 36 (2006): 18–39.

Pettit, Philip, and Michael Smith. "Freedom in Belief and Desire." *Journal of Philosophy* 93 (1996): 429–449; reprinted in Frank Jackson, Philip Pettit, and Michael Smith. *Mind, Morality and Explanation*. Oxford: Oxford University Press, 2004.

———. (1993). "Practical Unreason." *Mind* 102 (1993): 53–80 reprinted in Frank Jackson, Philip Pettit, and Michael Smith. *Mind, Morality and Explanation*. Oxford: Oxford University Press, 2004.

Prescott-Couch, Alexander. "Williams and Nietzche on the Significance of History for Moral Philosophy." *Journal of Nietzsche Studies* 45 (2014): 147–168.

Railton, Peter. "Reliance, Trust, and Belief." *Inquiry* 57 (2014): 122–50.

Rorty, Richard. *Philosophy and the Mirror of Nature*. Oxford: Basil Blackwell, 1980.

Rousseau, Jean-Jacques. *The Discourses and Other Early Political Writings*. Edited by Victor Gourevitch. Cambridge: Cambridge University Press, 1997.

Scanlon, T. M. *Being Realistic about Reasons*. Oxford: Oxford University Press, 2015.

Scott-Phillips, Thom. *Speaking Our Minds: Why Human Communication Is Different, and How Language Evolved to Make It Special*. London: Palgrave Macmillan, 2015.

Searle, John. *Making the Social World: The Structure of Human Civilization*. Oxford: Oxford University Press, 2015.

Sellars, Wilfred. *Empiricism and the Philosophy of Mind*. Cambridge, MA: Harvard University Press, 1997.

Shapiro, Scott J. *Legality*. Cambridge, MA: Harvard University Press, 2011.

Smith, Michael. *The Moral Problem*. Oxford: Blackwell, 1994.

———. "Rational Capacities, or: How to Distinguish Recklessness, Weakness and Compulsion." In *Weakness of Will and Practical Irrationality*. Edited by S. Stroud and C. Tappolet. Oxford: Oxford University Press, 2003.

Sober, Elliott, and David Sloan Wilson. *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press, 1998.

Sperber, Dan, and Deirdre Wilson. *Relevance: Communication and Cognition*. Oxford: Blackwell, 1986.

Stalnaker, Robert C. *Inquiry*. Cambridge, MA: MIT Press, 1984.

———. "Assertion." In Stalnaker*, Context and Content* (Oxford, Oxford University Press, 1999) 78–95.

Sterelny, Kim. *The Evolved Apprentice: How Evolution Made Humans Unique*. Cambridge, MA: MIT Press, 2012.

Stevenson, Charles L. *Ethics and Language*. New Haven, CT: Yale University Press, 1944.

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

Tomasello, Michael. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press, 2014.

———. *Origins of Human Communication*. Cambridge, MA: MIT Press, 2008.

Tuomela, Raimo. *The Importance of Us*. Stanford, CA, Stanford University Press, 1995.

Vargas, Manuel. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press, 2013.

Wallace, R. Jay. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press, 1996.

Williams, Bernard. *Truth and Truthfulness*. Princeton, NJ: Princeton University Press, 2002.

Winch, Peter. *The Idea of a Social Science and Its Relation to Philosophy*. London: Routledge, 1963.