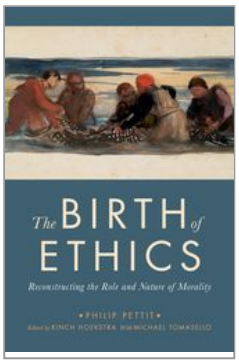

The Birth of Ethics: Reconstructing the Role and Nature of Morality
Philip Pettit and Kinch Hoekstra

Print publication date: 2018
Print ISBN-13: 9780190904913
Published to Oxford Scholarship Online: October 2018
DOI: 10.1093/oso/9780190904913.001.0001

Discovering Responsibility

Philip Pettit

DOI:10.1093/oso/9780190904913.003.0008

Abstract and Keywords

In the ordinary world, we identify agents whom we blame or hold responsible for a misdeed such as telling a lie, not just by the fact that it was possible for them to have spoken truly—this, in the absence of exemption or excuse—but also by three other presumptive facts. First, that they had a robust capacity to have told the truth; second, that they acted as they did in a context in which we think it was appropriate for anyone, including themselves, to exhort them to tell the truth; and third, that they are suitably reprimanded for not having acted as they were able and exhortable to act. Assume that in Erewhon certain standards get identified as multilaterally desirable by all lights, and as being within the reach of our capacity and exposed to the effect of exhortation. We will each be virtually pledged to those standards, not rejecting them in the manifest presence of an expectation that we will conform. And that means that, consistently with regarding you as someone whose pledges I can generally rely on, I must think that when you lie in the absence of exemption or excuse, you had the robust capacity to tell the truth, you were suitably exhortable to tell the truth, and you are appropriately reprimanded for failing to exercise that capacity. Unlike standard naturalistic alternatives, the emerging view of what makes you fit to be held responsible targets, not your indifference to others, but your failure to have exercised the capacity to tell the truth; but it does this without suggesting that you had a contra-causal form of freewill.

Keywords: responsible, blame, standards, capacity, exhortation

The observations made so far provide solid ground for thinking that we in Erewhon would evolve a conception of multilateral desirability akin to the established category of moral desirability. We would inevitably develop practices of avowing and pledging, co-avowing and co-pledging our attitudes. Those practices would make the introduction of various concepts of desirability more or less inevitable. And, confronted with conflicts between those notions, we would be pushed to develop the concept of what is multilaterally desirable: that is, desirable in a way that transcends the perspectives of different persons. This, plausibly, converges on the concept of the morally desirable that figures prominently in common usage.

If it is to give us a potential explanation of the emergence of ethics, however, the narrative must also explain how we in Erewhon can come to think in terms of responsibility as well as desirability. It must explain why we who have evolved the concept of the morally or multilaterally desirable would go on to hold one another responsible to certain judgments and standards of moral desirability. It must offer an account of why we would censure those who flout such moral judgments or standards and commend those who conform.

As argued in the last chapter, the concept of multilateral or moral desirability will apply with a particular salience in Erewhon to certain shared and routine standards, like those exemplified in social norms against lying, violence, and fraud, and against infidelity in avowals or pledges. Such standards will constitute properly moral norms: that is, norms that presuppose the availability of the concept of moral desirability among practitioners.

(p.198) The account to be given of why we would hold one another responsible to moral norms of this kind might also be invoked to make sense of how we could hold one another responsible to certain social norms or standards, whether of prudence or patriotism, law or epistemology. But the focus here will be on holding one another responsible for conforming to moral norms in particular. The focus will be entirely on moral responsibility, as it may be put, where this is understood as responsibility to moral standards, not as responsibility of any inherently special, moral character; responsibility, so it is assumed, amounts to more or less the same thing, regardless of the standards involved.

This chapter takes up the responsibility challenge in three sections. The first section explores what is involved in holding someone responsible for an action, in particular an action that breaches moral norms, distinguishing between three messages that are conveyed by holding the person responsible. While that first section analyzes how we in the ordinary world hold one another responsible, the second section looks at why we in Erewhon are likely to hold one another responsible to moral norms—shared and routine moral standards—after the same fashion: specifically, in a way that conveys the same three messages. The

final section turns to general issues about responsibility and freewill, showing how the theory supported by the narrative about Erewhon relates to contemporary rivals.

6.1 Responsibility characterized

Setting the scene

As in the case of desirability, it is essential in pushing forward the narrative 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 be successful. There are various accounts in the literature of what it means to hold someone responsible for having done something, and there will be some mention of **(p.199)** these in the final section. But rather than going into the debate between different approaches at this point, the line taken here will be to begin from an account that has two virtues, one substantive, the other methodological. On the substantive side, the account satisfies many of the common connotations of saying that someone is fit to be held responsible for an action. And on the methodological, 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 in which I hold you responsible for something I see as an undesirable choice—an offense or misdeed, as we may now say—and blame you for what you did. 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 fitness to be held responsible is likely to be at its clearest there. And if it proves possible to articulate the concept of responsibility for this scenario, then the lessons should carry over to the case in which I hold you responsible for having done something good rather than something bad.[1]

Fitness to be held responsible as conceptualized here is a role property that consists in your being such that, under the rules of the practice, you are an appropriate target for being held responsible. Fitness to be held responsible consists in your satisfying the messages or thoughts I convey about you when I hold you responsible. It may be realized in different individuals on the basis of varying neural configurations, as indeed it may be realized in different corporate bodies on varying organizational bases (Pettit 2007c). But the idea of fitness to be held **(p.200)** responsible can be defined in a way that abstracts from how it is realized or grounded in such different agents.[2]

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 might have made it desirable to hide the truth. And let the context of action be

one in which telling the truth, by common assumption, does not require heroic virtue but only a routine form of sensitivity to the desirability of truth-telling.

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 by way of holding you responsible for a misdeed when I say: “You could have done otherwise; you could have told the truth.” To hold you responsible, of course, is not necessarily to make any such utterance. But it is actively to assume an attitude that you could use those words or similar expressions to convey. It is actively to form a thought that you would be likely to communicate in that way, if it was possible to do so—for example, it did not concern someone in the distant past—and was not too difficult or costly (Watson 1987; McKenna 2012).

“You could have done otherwise”: the literal content

The literal meaning of “You could have done otherwise,” naturally understood, is just that you might have done otherwise: that it was possible, however unlikely, that you should have responded to the desirability of telling truth; there was nothing that stood in your way. But is the possibility of having done otherwise always going to obtain when **(p.201)** I think you are blameworthy? Yes, it is, contrary to a familiar line of thought (Frankfurt 1969).

Suppose we blame you for misleading me by telling a lie or, assuming this conveys the same message, by staying silent. And suppose we continue to blame you, after learning that you could not have done otherwise than mislead me; this, because, unbeknown to you, there was someone on standby, ready to intervene by preventing you from telling the truth. Blaming you in such a situation seems reasonable. So, does it show that we may reasonably blame you for doing something, even when it is not the case that you could have done otherwise?

No, it doesn’t. You could still have tried to tell the truth rather than misleading me spontaneously; you could have taken steps that required the standby agent to intervene. And in at least that sense, if not in the sense of actually telling the truth, you could have done otherwise than you did.

You could not have done otherwise even in that sense, of course, if, as is sometimes supposed, the standby intervener would have stopped you, not just from telling the truth, but from trying to tell the truth. But if you could not even have tried to tell the truth, you would not have enjoyed agency in the proper sense and I would not have grounds for blaming you; you would be wholly exempt, by ordinary criteria, from censure.

This shows that with any misdeed for which I hold you responsible, there must be some sense in which you could have done otherwise, some sense in which that possibility holds. And that being so, we may set aside the complexities of

the sort of case in which doing otherwise means just trying to do otherwise. We may stick with the standard sort of example in which I blame you for misleading me by saying, correctly, that it was possible for you to have told the truth.

What sort of possibility should we take this remark to ascribe? Not just logical possibility, for sure, but something like physical, psychological, and social possibility: in short, possibility of an agential kind. After all, we would not think that you could have done otherwise under the construal relevant to holding you responsible for an offense, if there were any physical, psychological, or social obstacles that got in your **(p.202)** way. The agential obstacles that are generally taken to undercut the point of saying that you could have done otherwise in such a context are related to exemption and excuse in the senses introduced earlier. That you could have done otherwise, construed to fit that context, means in its literal signification nothing more or less than that exempting and excusing factors were absent.

That exempting factors were absent means that you were not subject to any general agency-debilitating condition like paranoia or obsession or delusion or something of that kind. On the assumption that these come only in non-partial forms, which is maintained for convenience throughout this study, any such condition would remove altogether the possibility of your having done otherwise and exempt you from being held responsible (Watson 1987; Wallace 1996; Gardner 2007).

That excusing factors were absent—that is, unforeclosed, excusing factors—means that you were not confronted by specific, situational obstacles, epistemic or practical, that stopped you registering or acting on the desirability of truth-telling. On the assumption that these, too, come only in non-partial forms, any such factor would have blocked the possibility of your doing otherwise than telling a lie.

“You could have done otherwise”: three connotations

The literal content of saying that you could have done otherwise, by this account, is fairly platitudinous: that you were not subject to any exempting or excusing influences, so that there was nothing about your situation that ruled out acting other than how you did. Uttered by way of holding you responsible for a misdeed, however, I convey much more than this mere possibility in saying that you could have done otherwise: that you could have told the truth, responding to the desirability of that option (Nelkin 2011). As I communicate indirectly or pragmatically that I believe that p when I assert that p, so there are at least three messages about my attitude that I communicate in a pragmatic manner when I say that you could have done otherwise.

These messages are pragmatic because I convey them, not by virtue of the semantic content of the remark, but by virtue of the pragmatic **(p.203)**

significance of my uttering the sentence in the context of responding to a misdeed. In that sense, the messages are conversational implicatures (Grice 1975a). They are messages that I take to be manifestly available to you as audience insofar as you see me as a normal interlocutor and as someone, therefore, who has something more interesting to convey—something more relevant in the context (Sperber and Wilson 1986)—than the platitude associated with its semantic content. This is the platitude according to which you might have responded to the desirability of telling the truth in the sense that there was nothing like an exemption or excuse to stop you doing so.

Saying that you might have told the truth in this vein pragmatically conveys the following three attitudes over and beyond the belief that that possibility was open: first, a recognition that you had the capacity to tell the truth; second, a retrospective exhortation, to introduce a notion explained later, to have told the truth; and, third, a reprimand for not having told the truth. I recognize that despite not having acted like someone with a capacity or disposition to respond to the considerations that made truth-telling desirable, you did indeed have such a disposition at the time of choice. I exhort you to have told the truth, displaying an impatience at your failure; I hold by an attitude after the action that I might have expressed beforehand by saying that you can and ought to tell the truth, thereby urging you to do so. And finally, I reprimand you for not having told the truth; I present you as someone who is on the hook, without any exempting or excusing consideration to invoke.

The first connotation: recognizing capacity

If I say “You could have done otherwise” in response to a misdeed, then I credit you with a capacity or disposition to have done otherwise in the situation where you made your choice. I assume that you had that capacity and simply failed to exercise it. I do not conclude from your failure to act as appropriate that you must have lacked the capacity to act appropriately. Rather, I judge that you had that capacity but failed to manifest it in your behavior.

(p.204) Does the capacity I credit you with in saying that you could have done otherwise consist just in the fact that you were not subject to any physical, psychological, or social obstacle, of the kind that would exempt you or excuse you from being held responsible? No, it does not. The claim that you were not disabled in the excused or exempted fashion—that it was agentially possible for you to have done otherwise—is involved in the literal content of the words: You could have done otherwise. But the first connotation of using those words in holding you responsible is that you had the capacity to have done otherwise in a much more demanding sense than this.

It may be agentially possible for you to give away all your worldly goods, surplus to strict need, to those who are poorer than yourself. But I would hardly hold you responsible for not doing so. There is a sense of capacity in which we take this to

be beyond your capacity, so that if you did it, and I approved, then I would regard it as an exercise of heroic virtue. When I hold you responsible for failing to do something, however, I must think that there is a more substantive sense in which it lay within your capacity. Thus, when I hold you responsible for telling a lie by saying that you could have told the truth, I must take those words to convey something more than the fact that it was a bare, agential possibility that you might have told the truth: that there was no exempting or excusing factor to stop you.

Intuitively, I must convey the message that you were such that it was not only possible for you to register and act on the desirability of telling the truth; it was also to be expected that you would do so. Your makeup and milieu did not rule out telling the truth but, much more importantly, they actually ruled it in: they made it into a matter of natural expectation. You may not have told the truth in the actual situation, then, but the implicature is that you would have told the truth across many of the variations on that situation in which the considerations of desirability that support telling the truth remain saliently in place, and exempting and excusing factors continue to be absent. You had the capacity to tell the truth in that demanding sense of capacity, albeit you failed to exercise it.

(p.205) The message, in other words, is that it was not barely possible for you to respond to the desirability of truth-telling and speak the truth, as the absence of exempting and excusing factors would imply. It was dependably or robustly possible for you to do so. You were disposed to tell the truth robustly over variations of circumstance that continued to keep that desirability in place, and did not introduce excusing or exempting circumstances; in particular, you were disposed to tell the truth over variations in which different sorts of distractions or disturbances were in play. You were disposed to tell the truth so robustly, indeed, that your failure to do so must count as something of a fluke (Smith 2003; Hieronymi 2007; McGeer and Pettit 2015).

The message holds, it appears, even in the case in which you are a stranger, or are someone who lied to me in the past. If I am to hold you responsible for lying to me now, then, regardless of your missing or miserable record, I must credit you with a capacity to have told the truth. Otherwise, I would have to think that you were not properly responsive to the accepted desirability of truth-telling. And in that case, I could hardly hold you to the truth-telling standard, and blame you for not living up to it.

You might be someone, it is true, who is unresponsive to the accepted desirability of truth-telling. In that extreme case, I might blame you in the absence of exempting or excusing factors for having allowed yourself to become unresponsive or for not doing anything to regain responsiveness. But given that you are now someone incapable of robustly telling the truth, it would be misleading of me to blame you for the particular lie you told, as if that were

something out of the ordinary. What I would do, rather, is to excommunicate you, as we might say, from the moral community: to denounce you as someone wholly insensitive in the relevant, truth-telling domain.

As exemption and excuse exculpate you in one sense, excommunication would exculpate you in another. But the sort of exculpation it provides is even worse than blame or censure: it represents the sort of reaction that is appropriate for the irredeemably evil person. To blame you for telling a lie is to display a positive attitude of respect insofar as **(p.206)** it involves crediting you with a robust capacity to tell the truth, albeit a capacity you did not exercise on the occasion in question. To excommunicate or denounce you is to put you beyond even the pale of routine culpability; it is to cast you in the very lowest regions of Dante's hell.

That we assume a rich capacity to have done otherwise in anyone we blame for not living up to a shared, routine standard—and so a standard to which they themselves subscribe—is supported by the fact that, absent any excusing or exempting difficulty, such an offender is unlikely to claim that it would have required heroic virtue on their part to abide by the standard. They are unlikely to avoid regular blame or censure, in other words, when that means inviting excommunication instead. And that is to say that they may be expected to lay claim implicitly to having the robust capacity to have done otherwise—in our example, to have told the truth. If we make a default assumption that an unexcused, unexempted offender had the robust capacity to have told the truth, then, that only reflects an implicit invitation on their part to be ascribed such a capacity.

Still, it may seem rash of me to assume in every case in which excuse and exemption are unavailable that, despite having told a lie, you had the robust, situation-specific capacity to tell the truth. It may seem reckless to make the default assumption that you were sufficiently responsive to the desirability of truth-telling to make it surprising that you should have lied. In mitigation of this difficulty, however, there are three softening qualifications that may help to give the assumption some plausibility.

A first is that, while it is convenient to speak of the presence of such a capacity as an on-off matter—and while this way of speaking will be maintained here—it may be somewhat misleading. Having the required capacity does not mean responding invariably to the relevant considerations, at least in the absence of excuse and exemption. It means responding more or less dependably to those considerations, where the threshold at which the response counts as suitably dependable need not be set at the highest level.[3]

(p.207) A second softening qualification is that when I take you to have the capacity to respond to the desirability of truth-telling, and to be able to avoid

lying, I do not have to think that it will be easy for you to respond in that way. I may think that you can respond appropriately, not because doing so comes naturally to you, but because you can get yourself to respond in that way. You can take measures to guard, for example, against being distracted from the desideratum of telling the truth or against any disturbance of the motivating effect of that desideratum. For example, you can remind yourself of the reputational costs of failure, letting this reinforce the attraction of truth-telling and strengthen your will. Your capacity to respond to the desirability of truth-telling in a case in which you actually lie would then consist in the disposition to be moved in part by such costs, and so to tell the truth, in relevant variations on the actual circumstances.[4]

A third softening qualification of the capacity I credit you with when I say that you could have told the truth will be relevant in later discussion. I need not think of this as a capacity that you had on an enduring basis, independently of the effect I attribute to my having encouraged you to speak the truth—or to someone else's having done so—or independently of the fact that you had been reminded about the costs that lying would impose on a third party. I may think, in other words, that, while you had a dependable disposition at the point of action to tell the truth—a disposition that was partly grounded in the encouragement you received or of the reminder you had been given—that disposition was not of the standard, durable type; it was not necessarily a disposition that was already in place, prior to the effect of those pressures.[5]

(p.208) The second connotation: exhorting the agent

Saying that you could have told the truth in censuring you for lying implies, as mentioned, that it was something of a fluke that you failed to exercise that capacity. But in saying that you could have done otherwise, I do not mean to communicate just that it was a fluke that you did not tell the truth. That would be to console you rather than to speak in censure. Blame or censure involves quite a different attitude toward your failure to exercise your truth-telling capacity: a form of impatience with your failure; a refusal to accept it as a brute, regrettable fact. And the second connotation of saying that you could have done otherwise in such a case is to communicate that distinctive attitude.[6]

How to describe the attitude involved? Perhaps the best way to approach the issue is to consider the attitude that I or anybody else takes toward you prior to action, when we enjoin you to tell the truth, perhaps spelling that out by saying in an encouraging way: you can tell the truth. Such an injunction or exhortation says or supposes that you ought to tell the truth or that there is reason for you to tell the truth. But it does not have merely the force of a positive evaluation of that action. It does not amount just to saying that it would be better if you told the truth, where that has no action-directive significance for you now: i.e., where it has the same merely evaluative force as saying that it would be better if you had learned the local language or had received a scientific education. When the

words “You ought to tell the truth” constitute a mere evaluation, it makes sense to add: “but I realize that you won’t.” Where they constitute an exhortation or injunction, such an addition would make no sense; it would undermine the force of the utterance.

The attitude I take toward you in the wake of your failure to tell the truth is best cast as a counterpart of this exhortatory attitude. When **(p.209)** I say that you could have told the truth, I convey the message that it would have been appropriate for me or anyone else to have exhorted you to tell the truth prior to your utterance; it would have been appropriate, indeed, for you to have exhorted yourself to do so, seeking to get yourself to tell the truth. I retrospectively ratify that exhortatory attitude, implying that at the time of action you were a fit target for exhortation. You had a capacity to tell the truth that was sufficiently robust to have made sense of anyone’s exhortatory efforts to get you to tell the truth. You were suitably exhortable in the domain of the action you took.

In communicating this message, I refuse to be resigned to the negatively evaluated fact that you hid the truth, as I might be resigned to the weather having turned foul or the dog having misbehaved. I invite you to look at what you did, as we might say, and to recognize with me that it was within your power to have done otherwise; the suggestion is that this was not just a possibility, even a robust possibility, as it was just a possibility that the weather might stay fine or the dog mind its manners. The message is that whether you were to reveal or hide the truth was up to you—it was a matter of your choosing—and you blew it; you made a choice contrary to what I or you or anyone else might have reasonably exhorted you to do in advance.

This message may be described as one of retrospective exhortation, where that is understood as a message of ratifying the exhortatory efforts that someone might have made prior to the action. In retrospective exhortation, I express the sort of attitude toward what you have done that I would have expressed prior to the action if I had said in the normal exhortatory way, “You can do it—you can tell the truth,” or if I had implied this by enjoining you to do it. Speaking in retrospect, I cannot aspire to get you to change what you did, of course, but I can communicate that you did it in the presence of resources that would have made you a fit target of exhortation before the action. Fitness to be held responsible for an offence ex post goes with fitness to be exhorted not to do it ex ante.

(p.210) The third connotation: reprimanding the agent

The third effect of saying that you could have done otherwise in the case of a misdeed is to reprimand you: to communicate an unwelcome attribution of failure. Not only do I recognize your capacity to have responded to considerations of desirability and told the truth, treating this as a robust possibility. And not only do I display an exhortatory attitude, expressing

impatience rather than resignation at how you behave. 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, penalizing manner and thereby reprimand you for not having told the truth.

In communicating this reprimand, I present it as deserved in a straightforward sense, if not deserved in every sense possible (Pereboom 2014). I communicate as part of the implicature that you were not subject to any exempting or excusing obstacle that might let you off the hook and that you were not lacking in sensitivity to relevant considerations of desirability. And I convey thereby that you have no basis for complaining about the reprimand. In presenting that reprimand, of course, I may also support further sanctions of custom or law, whether these are justified on consequentialist, contractualist, or other grounds (Scanlon 1998, Ch. 5). But we may put aside penal sanctions here, concentrating only on the sort of reprimand or censure that they normally presuppose.

The task ahead

As remarked earlier, I express and communicate my belief that you could have told the truth—that there was nothing relevant to stop you doing so—when I report that possibility. By the implicatures just examined, however, I also convey a great deal more about my attitudes when I make this remark in the wake of an offense. As the words express a belief that you might have told the truth in such a context, so they will also express a recognition of your dispositional capacity to have told the truth, a retrospective exhortation to have told the truth—that is, a retrospective ratification of the exhortatory attitude—and a reprimand for **(p. 211)** having actually told a lie. They represent you as someone with a rich capacity to have told the truth, someone who is suitably exhortable on that front, and someone who is subject to an unwelcome ascription of failure that they have no grounds to complain about.

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 moral desirability. There is good reason, it turns out, 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 described.

As the discussion has focused on cases in which I hold you responsible for misdeeds rather than for good deeds, so it has focused on cases in which I hold you responsible, not on cases in which I hold myself responsible. But the points made about those interpersonal cases apply straightforwardly in the intrapersonal cases, as well. Insofar as it makes sense for me to recognize your capacity to have told the truth, to exhort you retrospectively to have told the truth and to reprimand you for actually having told a lie, so the same exercise

can make sense in relation to myself. I can remonstrate with myself as I can remonstrate with you about any failure, recognizing my own capacity in the relevant domain, exhorting myself to have exercised it, and reprimanding myself for not having done so. The narrative will not address this first-personal case directly, but it supports the same lessons there as it supports in the case of holding another person responsible.

Responsibility and regulation

Before developing the narrative to take account of responsibility, it is worth observing that the practice of holding one another responsible is very different from the blind regulation that was discussed in chapter 2 in the account of pre-social norms. The observation applies to the practice of holding one another responsible to any suitable standards, such as social norms, not just with the practice of holding one another **(p.212)** responsible to moral standards. But we concentrate here on the moral practice.

Blind regulation would consist in the fact that if you do not tell me the truth in any instance, that is likely to give me a bad opinion of your reliability as an informant, to lead me to speak ill of you, and to impose a reputational cost. And the prospect of such a cost is likely to lead you, instance by instance, to tell the truth in making reports and to conform to similarly welcome patterns like abstaining from violence or coercion or theft.

In the scenario envisaged with a pre-social norm like telling the truth, we each have an interest in proving ourselves to be reliable 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 pass 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 telling the truth and may be expected to elicit thereby a general pattern of truth-telling.

The practice of holding one another responsible for doing something like telling the truth will certainly have a similar, regulative rationale. In recognizing your capacity to have done otherwise than tell a lie, in exhorting you retrospectively to have told the truth, and in reprimanding you for not having done this, it should be clear in ordinary practice 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 intention may be just to draw your attention to the failure, presenting it as one you could and should have avoided. But there would scarcely be any point in holding you responsible for any failure of this kind, if I thought that there was no possibility of getting you to change. This is an observation supported even in a tradition

that insists on the distinction between practices of holding others responsible and efforts to reform them. As P.F. Strawson (1962, 25) says, if “our beliefs about the efficacy of some of these practices turn out to be false, then we may **(p.213)** have good reason for modifying or dropping some of those practices” (McGeer 2014).

But notwithstanding this regulative rationale, the practice of holding you responsible for telling the truth is distinct from the reputational regulation exemplified in the pre-social case. As emphasized in discussing that case, I and others may exercise such regulation without being aware of the general pattern that we thereby elicit in our aggregate behavior toward one another. And so, it may be an exercise in which none of us intentionally takes part. Not recognizing what I am doing in regulating you, I cannot do it intentionally; it is not something I desire as such, for example, nor is it a foreseeable effect of something that I desire as such.

When I hold you responsible for living up to a standard like truth-telling, I do so with an awareness of the overall pattern associated with the standard and out of a desire that mere regulation need not involve. Thus, if I censure you for not acting as the desirability of truth-telling requires, I censure you consciously and intentionally. This will certainly be so if I express the censure in words, as in saying “You could have done otherwise,” “You could have told the truth.” But it will also be the case if I just assume an attitude of censure, in which assuming an attitude is a mental act. It barely makes sense—although it may convey something metaphorically—to imagine that I might censure you but only unconsciously or unintentionally.[7]

But while regulating one another reputationally into certain patterns of behavior falls short of holding one another responsible to corresponding standards, that regulative regime may continue to play an important role in supporting a 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 **(p.214)** reputational discipline in which, as a matter of common awareness, we expect one another to be suitably reliable: in particular, reliable in telling the truth. The existence of this background pressure is bound to increase the prospects for our regulative success in holding one another responsible for telling the truth and for conforming in other ways to the moral standards that we take for granted.

This observation connects with a point registered at different points in earlier discussion. The motivating engine that drives us to conform to many moral or indeed social norms may continue to be substantively reputational, even as we steer by a distinct justifying rationale for conformity. The steering thought in the social case is that conformity is a precondition for communal acceptance, in the moral that it is required as a matter of moral desirability. In holding one another

to moral norms, as in holding one another to social norms, it is plausible that we expect the rationale invoked to be effective in good part by virtue of the reputational effects that are going to be salient on all sides.

6.2 Recognition, exhortation, reprimand

Holding one another responsible in Erewhon

Why might we be led in Erewhon to go beyond mere regulation and to hold one another responsible for living up to certain moral standards, in particular standards that are shared and routine? In order to answer this question, it will be enough to show that if in response to breach of a norm we make a remark like “You could have done otherwise”—the literal content of the remark will allow us to do this in the absence of excuse and exemption—then that will have the three connotations or effects associated with holding you responsible. If the remark reliably has such effects when made in response to a breach, then it must constitute an act of holding you responsible for the offense.

The three connotations may be described as the recognition effect, the exhortation effect, and the reprimand effect. The discussion that follows explores each of these in turn. While they may be triggered by **(p.215)** a breach of any shared and routine standards, the focus will be on the breach only of moral norms: that is, the shared and routine standards of moral desirability that get to be established among us in Erewhon.

The recognition effect

According to the first connotation, the remark, “You could have done otherwise”, when uttered in response to a moral offense, does not merely record the fact that you might have done otherwise, given the absence of excuse or exemption. It conveys the message that you were disposed in the situation of choice to respond robustly to the relevant considerations of moral desirability: that is, to register and act on them in most variations on the actual circumstances—in particular, ones introducing various distractors or disturbers—in which the considerations remained present and there was no exemption or excuse. And so, it would communicate that your failure to respond appropriately in the actual situation was more or less a fluke; it was out of character.

But why would the remark be expected to attract this reading among us? Why would I not be moved by the evidence of your failure to conclude that, actually, you were not suitably responsive to our shared moral standards? It was possible, to be sure, that you might have responded to the standards, given the absence of exemption and excuse. But why should I think that this was a suitably robust possibility; why should I think that you would have responded to the standards in a wide raft of variations on the actual circumstances? Why should I ascribe a capacity to respond to those standards and conclude merely that you failed to exercise it? Why should I treat your behavior as a departure from normal: a contingent failure?

It would certainly be reasonable to treat it as a contingent failure if you had already demonstrated that capacity over a range of similar cases. But even in the absence of demonstrated capacity, as appeared earlier, we in the ordinary world assume as a default that you have such a capacity when we hold you responsible to shared, routine standards like those that support telling the truth. Indeed, we make this default assumption with others in general, even with strangers and even with **(p.216)** those who have lied to us in the past. So, is it possible to explain why we who live in Erewhon might support a default assumption of this kind?

Plausibly, it is. In order to provide the required explanation, however, it is essential to develop our earlier account of pledging and to defend a general thesis that applies in any instance of pledging. This will be described here as the anti-expulsion thesis.[8]

According to the anti-expulsion thesis, if pledging is to have a useful place in Erewhon, then we must generally take one another to have the capacity to live up to our pledges. In particular, we must not expel someone from the ranks of those who can live up to their pledges just because they fail on a particular occasion to do so. We may become more wary about assuming the ability to live up to their pledges in those who fail time after time, eventually excluding them altogether from the ranks of those with whom we can do business. But we should be slow to resort to such expulsion.

The reason for this is grounded in our nature as beings who need to be able to rely on one another in various ways, if we are to prosper and thrive. If we were disposed to expel one another from the ranks of the potentially reliable in response to any single failure to prove reliable, or even to any handful of failures, then we would be likely to deprive ourselves of the benefits of mutual reliance.

Assuming that everyone is going to fail occasionally to live up to a pledge, a disposition to expel someone on the basis of a single offense would eventually lead each of us to expel most others from the circle of the potentially reliable. And as we would each expel others from that circle, so we would each be expelled ourselves. As a community we would be led inevitably toward a social wasteland. It might not involve the war of all against all that Hobbes (1994, 13.9) imagined, but it would still be likely, in his words, to make life solitary, poor, nasty, brutish, and short.

(p.217) Back now to the question raised. Why might we make the default assumption in Erewhon that when you offend against a shared, routine standard of desirability, and there is no exemption or unforeclosed excuse available, the offense is not due to a lack of the capacity to respond to such a standard but just to a failure to have exercised the capacity? The answer can be set out in a

schematic argument, which relies for its conclusion on the anti-expulsion thesis just defended.

The argument goes as follows:

- We, your compatriots, will manifestly expect you, as we will expect others in general, to acknowledge and abide by the unquestioned requirements of shared, routine standards of desirability, at least in the absence of exemption or of unforeclosed excuses; that is what makes the standards shared and routine.
- We will take you in effect to have pledged fidelity to any such standard for, as we see things, you will have made the pledge in a virtual manner by not saying “Nay” to that manifest expectation—that is, by not denying that the standard is shared or routine.
- I would be refusing you the status of a potentially reliable pledger, if I were to conclude from one failure, or even a few failures, that you did not have the capacity to prove faithful.
- But it would be unappealing to refuse you this status in view of the anti-expulsion thesis; expelling you from the community of those capable of living up to pledges would rule out the prospect of future benefits and, as a general strategy, it would be destructive of mutual reliance.
- Therefore, in the case of shared, routine standards of moral desirability, we in Erewhon must be expected to make a default assumption that in any non-exempt, non-excused offense, you possessed but failed to exercise a robust capacity to conform.

This argument shows that, absent exemption or excuse, I am likely to respond to a failure on your part to conform to a shared, routine standard of truth-telling, not by questioning your capacity to conform, but by insisting that you had that capacity and that you failed to exercise it. And I am likely, moreover, to give this response a default status, **(p.218)** displaying it when I meet you for the first time, for example, or even when you have lied to me on previous occasions. I might be driven to withdraw the assumption, of course—I might be forced to treat you as an unredeemable liar, for example—if your failure was repeated time and time again. But the cost of expelling you in this way from my circle would be enormous, and so I would be likely to embrace it only as a last resort. It would amount to excommunication, as that was described earlier.

To resort to excommunication would be to treat you in effect as unconversable, or at least unconversable in the domain of the failure (Pettit and Smith 1996). For it would mean that I could not take you to be someone with whom it is possible to interact normally in that domain. I would recognize no basis for relying on your words, no basis for entering into a relationship where your

reports, avowals, and pledges guide me in my actions and expectations in relation to you.

The situation would be close to that in which I take you to be exempt from responsibility and unfit to be treated as a proper interlocutor. I might think of trying to engineer this or that response in you—I might deal with you as a subject for treatment (Strawson 1962)—but I could not seek to induce the response by pointing to its desirability under a routine, shared standard. In what follows, the assumption will be that we rarely resort to excommunication in Erewhon and that this makes sense; we are rarely confronted with the sort of unresponsiveness to shared, routine standards that would warrant excommunication.

The exhortation effect

If these considerations are sound, then when I say that you could have done otherwise in response to a misdeed, in particular some misdeed in which you were not subject to a suitable exempting or excusing condition, then I should be taken to convey by those words that you had the capacity to do otherwise. You were someone disposed to register considerations of desirability robustly—considerations derived, within the focus adopted here, from shared, routine standards of moral **(p.219)** desirability—and to act robustly as they require: robustly, in particular, over variations in distraction and disturbance.

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. "You could have told the truth," for example, should communicate a retrospective exhortation or injunction to have responded as the considerations of desirability required, not a resigned acceptance of a failure that I evaluate negatively. It should retrospectively ratify the exhortatory efforts I or anybody else—or you yourself—might have made to get you to tell the truth. It should express the same attitude that I might have expressed prior to the action by saying in the normal, exhortatory way: "You can do it: you can tell the truth."

Is it possible to explain why in Erewhon the words would communicate this second message? Yes, it is. Suppose 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 shared, routine 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: that they would express the sort of attitude that a prospective exhortation would have conveyed? For reasons related to the reputational discipline that operates there, it turns out that there is.

By the argument presented in support of the recognition effect, we in Erewhon are each pledged to live up to shared, routine standards of desirability like that of truth-telling. As assumed, these standards unquestionably require us, now here, now there, to act appropriately and to live up to them in that sense. We are pledged to live up to that sort of standard in the same way that we are pledged to act on any more specific intention to which we may have expressly committed ourselves. But just as reflecting on something that holds of pledging in general helped to explain why saying “You could have done otherwise” has the recognition effect, so reflecting on another general feature of pledging can help to explain why those words have the exhortation effect, too.

As the anti-expulsion thesis helped in the previous case, so a similar lesson—the empowerment thesis, as it will be dubbed—can help in the **(p.220)** present instance. The thesis is that we each empower one another in the practice of pledging. We help to elicit a capacity in one another to live up to our pledges and to get us each to exercise that capacity as appropriate.

Consider a situation in which you pledge to do something, inviting me implicitly to rely on your doing it. In any such case, it is manifest that you will be likely to suffer a serious reputational cost if you fail to perform as pledged: a cost in my opinion of you and in the opinion of any others who witness or learn of the failure. In the event of failure, we may not expel you from the community of the potentially reliable, but we will call you out as someone who proved unreliable in practice, someone who failed to exercise a robust capacity to prove reliable. The pledge means that you will not be in a position to excuse your failure by appeal either to a misleading or to a changed mind. Thus, when you make a pledge, you must do so in awareness that the stakes are high and that you will almost certainly suffer ignominy in the event of failure.

How can you be confident enough to make such a pledge? By the analysis offered earlier, you may draw some confidence from the general fact that there are lots of desiderata that make it attractive to act as the pledge requires. But since those desiderata may not remain relevant at the time for action, you must draw confidence in particular from the fact that acting on the pledge—proving faithful to your word—has a powerful reputational payoff. It promises to establish with a special force that you can live up to your word, earning you a very attractive reputation with anyone who learns, even learns after the event, about your performance.

But not only is it the case that you can draw enough confidence from the reputational discipline under which you operate to be ready to pledge yourself in this or that respect. We others must also draw confidence from the role of that discipline in your psychology, if we are to be ready in general to rely on the pledges you make. For why else would we be prepared to rely on you? After all, it will be obvious to us, as it is obvious to you, that living up to the pledge may

prove independently very unattractive and that only the reputational discipline may be there to keep you in line.

(p.221) By the narrative developed so far, pledging emerges in Erewhon side by side with avowal and becomes a stable feature of the society. On this assumption, your confidence, and our confidence, that you can remain faithful to your pledges is vindicated. We are vindicated in the belief that by being there to observe how you perform in keeping your pledges, and to form a judgment about your reliability, we can help to establish your capacity to live up to those pledges and to elicit its regular exercise.

What we others do for you, of course, we all do for one another. And that is just to say that the empowerment thesis is sound. We play a powerful role in getting one another to establish and exercise the capacity to live up to our pledges. By being there to impose reputational costs on any failure, we help to police one another, if only in the fashion of blind regulation, into acting as we pledged ourselves to act. We provide an environment for one another that scaffolds our capacity to live up to our words. That capacity is grounded, not just in how we each are in ourselves, but in how it is with those around us; it has an ecological, and not a purely psychological character (McGeer 2013; McGeer and Pettit 2015).[9]

The empowerment thesis explains why exhortation or injunction can play an important role in interpersonal interaction. The soundness of the thesis is a matter of common accessibility and awareness: the evidence supporting it is salient; that the evidence is salient is itself salient; and so on. And that means that, being aware of my reputational, empowering role, I can actively build on that role by encouraging you to live up to a pledge. I can exhort or enjoin you to be faithful to the pledge, and do so with a manifest rationale. By encouraging you, I express the **(p.222)** expectations that I and presumably others hold and I make the reputational cost of a failure particularly salient.

Suppose you have pledged to join others in some venture, then, and that you now shrink from the prospect. You may have pledged to go on a hunt but recoil from doing so, as in an earlier example, because the weather has turned foul. I may insist that you still have the capacity not to backslide, saying, “You can do it; you can live up to your word.” Or I may presuppose the presence of that capacity, as in saying, “You ought to keep your word and join the hunt.” And by making such an utterance, I may expect to have some effect in reminding you of the costs at stake and in bolstering your capacity to do what I am saying or implying that you can do. I may expect to empower you, if you need empowerment, and to help to get you to the point where acting on the pledge comes within your reach.

This lesson about the role of exhortation carries over from ordinary pledges to virtual pledges to live by shared, routine standards of desirability. By the argument run in support of the recognition effect, you are pledged in this way to be faithful to such standards. And so, I or anyone else may sensibly exhort or enjoin you to live up to any shared, routine standard of desirability. I will do this whenever I tell you what you are now required to do under such a standard, or when I just say that you can live up to the current requirements of the standard. I will play a role in helping you to conform to the standard, whether or not my help is needed; my words will have the empowering force of an exhortation.

These observations make it possible, finally, to explain why in an appropriate context the remark “You could have done otherwise”—in our example, “You could have told the truth”—has the character of a retrospective exhortation. They make sense of why those words express the sort of attitude toward you that anyone would have expressed prior to the utterance, if they had exhorted you to tell the truth.

If my saying “You can tell the truth” has the force of an empowering exhortation when uttered prior to a choice, then saying “You could have told the truth” in the wake of the choice is going to communicate that the ex ante exhortation was appropriate; it is going to ratify that **(p.223)** exhortation ex post. The alternative to the ex post remark would be to say “It’s not the case that you could have told the truth,” which would certainly convey the message that the earlier exhortation had been inappropriate. If it is not the case that you could have told the truth at the time of choice, after all, then it was misconceived on my part to tell you, with whatever exhortatory or injunctive intent, that you could do so.

Why would I reject that alternative in the wake of the choice, then, and say that despite failing to do so, you could have told the truth? I say this, presumably, because I maintain the same attitude toward you that I expressed in the original exhortation. The remark can only communicate that it was appropriate on my part to have enjoined you earlier to tell the truth; telling the truth at that time was within your reach, at least in the presence of my exhortation.

If this is right, plausibly, then whenever I say that despite a failure, you could have told the truth, I communicate that an earlier exhortation to tell the truth would have been appropriate. That will be the case, presumably, whether or not I or anyone else actually exhorted you at that time to tell the truth. And so “You could have told the truth,” uttered in a suitable context, must have the force of a retrospective exhortation, as that is understood here. It does not just express the belief that you had the capacity to tell the truth: i.e., that it was possible for you, robustly over distraction and disturbance, to have told the truth. It conveys an impatience with your having failed to exercise the capacity, given the status you enjoy of someone embedded in our culture of mutual exhortation and regulation.

Insofar as we see you as someone party to that culture, who was susceptible to prospective exhortation to tell the truth, we see you as a person with whom it makes sense to ratify that exhortation retrospectively.

This lesson extends to every case in which you offend against a shared, routine moral standard. In any such event, my saying that you could have done otherwise will not just express a belief in your capacity to have done otherwise—this, in the robust, dispositional sense of capacity mobilized by the first connotation—but also an exhortatory form of impatience with your failure to have exercised that capacity. The words will identify you as one of our mutually regulating kind and **(p.224)** will convey an exhortation to have done better as surely as they will convey a recognition of your dispositional capacity to have reached that level of performance.[10]

The reprimand effect

The observations so far show that assuming that no one is excommunicated, I and you and others in Erewhon would routinely satisfy the first two conditions associated with holding someone responsible. I would be in a position to give default recognition to your capacity, absent exemption or unforeclosed excuse, to respond to the considerations of desirability that require you to tell the truth; and to do this, even in the wake of failure. And I would be in a position to speak with an exhortatory 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 reprimand you by making such a remark, expressing an unwelcome opinion of your performance and communicating at the same time that this opinion is deserved.

By the account developed so far, my saying you could have done otherwise in the wake of a misdeed like telling a lie presupposes that it was manifestly appropriate for anyone prior to your action to have exhorted **(p.225)** or enjoined you to respond to considerations of desirability and tell the truth: this, assuming that the option of telling the truth was supported by a shared, routine standard of desirability. But if it was manifestly appropriate for anyone to have enjoined you to respond to considerations 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, plausibly, I can be taken to register 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 exhortation or 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: i.e., your failure to do what it would have been morally desirable for you to

do. 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 deserved in a straightforward sense. In saying that you could have done otherwise, conveying the message that you acted against an appropriate injunction, I assume that you were not subject to an exempting or unforeclosed, excusing condition and that you were not lacking in sensitivity to the relevant considerations of moral desirability. And in assuming 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 deny you any basis on which to complain about my reprimand.

But will the reprimand count as deserved in this sense if there is a naturalistic explanation for your having failed to exercise a relevant, robust capacity, say, the capacity to tell the truth? Or will the availability of such an explanation support the claim that despite the absence of excusing or exempting factors, still it is not fair to blame you? It might seem so. *Tout comprendre c'est tout pardonner*, as it is said; to understand everything is to pardon everything. But however things seem, they are not so. Explanation does not undermine reprimand.

(p.226) One sort of explanation for a failure is readily squared with maintaining that a reprimand is deserved. This is the kind that invokes a familiar sort of distraction or disturbance to make sense of why you failed to tell the truth. Thus, I might explain your lying to me by your having been subject to one or another temptation—say, to impress an audience—and your having proved weak of will. And I might explain any of a range of similar failures by appeal to factors like inattention or laziness. These explanations are fully consistent with maintaining the reprimand, since they are factors that we will take you, like more or less anyone else, to be capable of overcoming. In particular, they are factors that it will make sense for us to exhort you to overcome, the assumption being that you are able to respond to the considerations we put before you in the course of exhortation.

But there is another sort of explanation that we must recognize for any failure on your part to exercise a capacity to tell the truth, whether or not this also involved a failure to overcome distraction or disturbance. On naturalistic assumptions of the kind maintained here, the failure must have been sourced in some natural antecedent, probabilistic or deterministic. Some breakdown of normal functioning—say, an unknown glitch or brute chance—must have led you to behave out of character and not exercise your capacity. Does the availability of this sort of explanation, which we may or may not have the neuroscientific knowledge to detail, suggest that a reprimand is out of place? No, it does not.

Under the practice documented in this narrative, we invite each other to rely on us to keep our pledges—and, in the relevant case, to live up to shared, routine standards of desirability—in the absence of exemption or excuse, assuming in

ourselves an appropriate sensitivity to considerations of desirability: the sensitivity presupposed in the notion of a robust capacity. And we do this with confidence, since it is a matter of common belief among us that the reputational force mobilized by the invitation can generally lead us to be able to keep the pledges we make; if this were not a matter of common belief, then our relying on one another's pledges would be inexplicable.

If the practice allowed us to get off the reputational hook, however, just by citing a non-exempting, non-excusing cause of the failure—just **(p.227)** by invoking an unknown glitch or brute chance—this common belief would break down. For then the cost attendant on a failure would always be avoidable and we would lose any reason to expect, or expect others to expect, that the reputational discipline would make us into reliable pledgers. To think that you could get out of a pledge—including a virtual pledge to live up to a shared, routine standard of desirability—by citing a non-exempting, non-excusing cause of failure would be to display a misunderstanding of how things work in the practice of holding one another responsible, as that has crystallized in the course of the narrative.

By the account supported in the narrative, there are two assumptions built into the practice. The first is that regardless of our capacities, we cannot expect to regulate one another into conformity with any pledge, and so into conformity with any shared, routine standard of desirability, if there are exempting factors or unforeclosed, excusing factors present. And the second is that in the absence of such factors, and regardless of whatever other causes may be at work, we can expect to be able to exercise a significant regulative influence on one another. The factors that count as exempting or excusing, and the factors that count as mere distractions or disturbances, may shift from time to time, and may differ from those recognized in other societies; different cultures of expectation and regulation may cast somewhat different factors in these roles. But it is crucial to the practice of holding one another responsible to accepted standards that some causes of failure are treated as regulation-resistant and others as regulation-susceptible; exempting and excusing causes are treated as resistant to the effects of regulation, other causes as susceptible to those effects.[11]

Why, then, should you be expected to pay the costs associated with a misdeed that is due to a regulation-susceptible factor—say, a brute **(p.228)** chance or an unknown glitch? Why should you be expected to treat a reprimand as deserved? In a word, because the glitch or chance counts as an influence that you are able, by received ideas, to overcome. You have all the motivation required to carry you past it, given the reputational force field in which you live. Factors that count as regulation-resistant, and that we regard as exempting or excusing, are influences that you are unable, despite your assumed sensitivity to considerations of desirability, to overcome in the same way. It is not because they obstruct the operation of an allegedly uncaused will, or anything of that non-naturalistic kind (more on this later) that they are special. They are special

because they stand out among natural causes by virtue of the fact that there is not much that you can do, and not much that we can get you to do—this, by issuing appropriate exhortations and reprimands—that would lead you to overcome them.

The regulative aspect of the practice of holding responsible connects with the empowerment thesis mentioned earlier and with the ecological character of the capacity I ascribe to you when I say, given the absence of exemption or unforeclosed excuses, that you could have responded to the desirability of truth-telling. And it connects also with the earlier observation that the capacity I ascribe to you in making such a remark may be a capacity that you have only as an addressee of the sorts of expectations that the remark expresses. The capacity may depend for its existence and effectiveness on the presence of a community in which we hold one another to a standard like truth-telling. The theme will return in the next section in the course of a comparison between the approach to responsibility supported here and other approaches in the literature.

Back to obligation

The upshot of these considerations is that when uttered in a suitable context, the assertion in Erewhon that you could have done otherwise is going to serve the three functions required under a practice of holding an offender responsible; it will support the recognition, exhortation, and reprimand effects. This concludes the narrative explanation of how, having developed the idea of desirability, we would go on in Erewhon to **(p.229)** hold one another responsible for living up to shared, routine standards of moral desirability.

But there is still one loose end to tie up. By most accounts, it is the concept of the obligatory that is central to ethics, not the concept of the morally desirable. And by most accounts, it is standards of obligation, not standards of desirability, that ought to figure in our practice of holding one another to account. So how does obligation fit into the picture?

On the line adopted earlier, it is obligatory for an agent to choose one option rather than another just when it is the morally most desirable alternative among ‘erogatory’ options: i.e. options that are undemanding enough not to count as supererogatory. Thus, with any choice that violates a shared, routine standard of desirability, we will be in a position to think of the option of living up to that standard—meeting its unquestioned requirements in this or that situation—as obligatory. For given the routine nature of the standard, living up to it will count by definition as a suitably undemanding option. And given the shared status of the standard, living up to it will be the morally desirable thing to do, by perceptions that we share with the agent.

With the concept of the morally obligatory in hand, as mentioned before, we will also have access to the concepts of the morally prohibited and the morally permitted. An option will be morally prohibited if it is morally obligatory to avoid it. And an option will be morally permitted on a first usage, if it is not morally obligatory to avoid it and, on a second, if in addition it is also not morally obligatory to take it. On the first usage, the fact that an option is permitted does not rule out its being obligatory; on the second, it does: the option, as it is said, is merely permitted.

However important the category of the obligatory is in Erewhon, there is no reason to think that it will take over completely within our discourse. Certain standards might be shared across the community, without being routine enough to count as obligatory. They might manifestly require such a level of effort and difficulty that we would not be prepared to blame people for failing to conform to them; they might be taken to fall outside people's robust capacity. Such standards would **(p.230)** count as supererogatory ideals, not matters of obligation. It is important to keep them in the picture if only because, as already noted, today's supererogatory ideals may become tomorrow's standards of obligation—this, as a result of a changing sense of moral desirability, and a changing pattern of mutual, empowering expectation.

6.3 Theories of responsibility

Two other approaches to responsibility

It may be useful in conclusion to try to relate the theory of responsibility supported by this narrative to more familiar accounts. The theory supported is naturalistic in the sense that it is compatible with the assumption that all the entities in the world, including human beings, belong to the domain charted in the most basic natural sciences and are subject to the laws established there. The narrative provides a presumptively naturalistic explanation of how we in Erewhon could have come to develop the concept of moral responsibility, as it provides a naturalistic explanation of how we could have come to develop the concept of moral desirability. And so, it does not take that concept to ascribe a non-naturalistic property; in particular, it does not take our fitness to be held responsible to presuppose a non-natural, scientifically alien freewill.

The theory supported in the preceding narrative contrasts in the first instance, then, with the family of non-naturalistic or libertarian views that posit such a freewill. A libertarian theory would represent the capacity to respond to standards of moral desirability as a sui generis capacity on your part to intrude yourself into the causal chain preceding any action—a causal chain involving the immediate neural antecedents of behavior—and to make your behavior conform to what those considerations support. It constitutes a capacity, on this sort of account, to inhibit the neural chain that might take you in an unwanted direction and so to control things according to your will. The capacity postulated is non-natural, for it implies that the causal network that operates at a **(p.231)** level

accessible to natural science does not constrain you in any choice, or even expose you to the vagaries of objective chance. It leaves you with leeway to act according to your own will and so, if that be your will, to act according to the perceived demands of moral desirability.

The presence of such a capacity, according to libertarian views, means that whatever you do, you do as a result of what you as a person—you as identified with your will—want to do, not as a result of causal laws and pressures that operate on you or within you. Hence the capacity makes sense of why we might say in advance of action that you should and can do something like tell the truth: you can exercise your will, notwithstanding any causal obstacles, so as to respond to the considerations that support that option.

That being so, libertarian views also make sense of why we might say in the wake of an offense that you could have done otherwise, conveying an injunction to have done otherwise, and not just a negative evaluation of the choice you made. The idea is that despite the fact that natural antecedents may have prompted the misdeed, you had the capacity to override those influences and act as the relevant standard requires. Thus, when we exhort you to have done otherwise, we seek to remind you of that capacity and of your failure to have exercised it.

No naturalistic theory can live with such an image of the person and so it has to offer a very different sort of account of what it is about you or anyone else that allows us to think that, absent exemption or unforeclosed excuse, you have or had the capacity to live up to standards of moral desirability, responding appropriately to their requirements. The account has to cohere with an image of the human being as a creature, like any other animal, that is organized out of cellular—and ultimately, molecular, atomic, and sub-atomic—matter; and a creature, therefore, that operates under the force of the laws that govern such matter at the most fundamental level. It has to make sense of your capacity to live up to moral standards on the assumption that its exercise does not involve a rupture in the regular causal order.

One candidate analysis, once popular among naturalists, is conditional in form. It holds that to have the capacity to respond appropriately to considerations of desirability is to be such that if you want **(p.232)** or try to respond to those considerations—if that causal antecedent is in place—then you will respond to them. But no such analysis can be satisfactory (Chisholm 1982). For it may be true that if you wanted or tried to respond to such considerations, you would respond appropriately, without its being true that you could want or try to do that. You may be a psychopath who satisfies the condition given but could never want or try to respond to relevant considerations. And in that case, we would scarcely credit you with the capacity to respond appropriately.

The theory of responsibility supported in the narrative of this chapter offers a naturalistic account that avoids that problem, while still making sense of what we do in holding someone responsible. But this theory should be distinguished, not only from the non-naturalistic family of approaches, but also from an alternative family of naturalistic views that also seek to avoid that problem (see Clarke, McKenna, and Smith 2015).

Reconstructing this family of approaches in a way that abstracts from different versions, they all suggest that to say you could have done otherwise in censuring a lie is to hold that you were not exempt from being held responsible—you were not subject to a debilitating disorder, for example, or compulsion—and that you could have had a better attitude, displaying it in telling me the truth (Wallace 1996). There may be no relevant sense in which you could have acted otherwise in the situation of choice, according to this sort of view; the general assumption is that you could have done otherwise in that sense only if you had a non-natural power of will. But there is a sense, so the idea goes, in which you could have displayed a better attitude. There is a possibility, perhaps even a robust possibility, that you might have become sensitized to the salient desirability of truth-telling, might have developed a suitable attitude toward truth-telling, and might have been led therefore to tell the truth in the situation on hand. You could have been otherwise, even if you could not have acted otherwise; you could have been generally good willed.

Proponents of this approach typically assume that the relevant attitude is characteristic of your makeup and properly attributable to you, an idea that was introduced by Gary Watson (1996); see, too, Wolf **(p.233)** (1990). Most now follow T.M. Scanlon (1998, 20) in assuming further that the characteristic, attributable attitude must be “judgment-sensitive,” that is, responsive to reasoning and such that we may expect it to be present in anyone who reasons properly, making appropriate judgments of desirability; see, for example, Smith (2005).

According to this approach, then, three facts make it appropriate to hold you responsible for telling a lie. First, that it was possible for you, perhaps even robustly possible for you, to have had a better attitude and, absent exempting or excusing obstacles, told the truth in the context in question. Second, that this possibility would have materialized if you had been generally more sensitive—say, more sensitive over the course of your life, or over some relevant period—to the considerations of desirability that supported truth-telling in appropriate contexts. And third, that there was nothing in your character or circumstances that put such sensitivity beyond your reach: it would have been available to you had you been conscientious about the meaning and implications of judgments of desirability that you were in position to make.

The crucial difference between this attitude-centered account, as it may be called, and the action-centered account supported here is that on the former account there is no sense in which it is appropriate to exhort you to have acted differently in the particular choice indicted. Why exhort you to have done something that was the product of naturalistic cause or chance, so the idea goes, and not something that you had the non-naturalistic power, regardless of such antecedents, to control? Despairing of finding sense in such a possibility, the focus in the approach is on the fact that you could have had a different sort of attitude and could have been a different sort of person: for example, you could have been someone of goodwill toward others. And so blame is associated with finding fault with you for the sort of attitude you displayed—for your lack of goodwill—rather than for the particular action you took (Scanlon 2008). The idea, presumably, is that the problem raised by the assumption that you could have done otherwise is not raised by the assumption that in the appropriate sense you could have had a better attitude and could have been a better person.

(p.234) Recasting the theory supported here

The theory supported by the narrative of this chapter focuses on the action that you performed rather than just on the sort of attitude you displayed and the sort of agent you proved yourself to be. But it holds that this focus is consistent with believing that you, like any other human being, are subject to the laws that govern the natural world. In particular, as emphasized earlier, it is consistent with believing that the action I censure you for performing was the product of a natural cause or natural chance. All that it presupposes is that the action was not occasioned by a factor of a suitably exempting or excusing kind.

The theory is that the capacity that makes you fit to be held responsible for a misdeed like telling a lie is a disposition on your part, in the absence of a suitable exemption or excuse, to respond robustly to relevant considerations and tell the truth. But it denies that when I say that you could have done otherwise and told the truth, the principal message is that it was a relative fluke that you told a lie. For the theory insists, at a second level, that in saying you could have done otherwise, I speak from the viewpoint of an exhortatory interlocutor, reaffirming the injunction to tell the truth that someone might have addressed to you before you spoke. And that means at a third level that the remark that you could have done otherwise constitutes a reprimand and, given the absence of exemption or excuse, a deserved reprimand that you are unable, under the relevant practice, to deflect.

This action-centered theory endorses a purely naturalistic metaphysics, making sense of how I can hold you responsible for a particular misdeed by appeal to the exhortatory or regulatory perspective I adopt in doing so. Your failure to live up to a standard like that of telling the truth will look from within the exhortatory perspective as an instance in which I or anyone else—or perhaps even you yourself—might well have gotten you to tell the truth by being there before the

choice to force your attention on the standard, to encourage you to conform to it, and to make manifest the cost of failing. When I say that you could have done otherwise, I stand by the appropriateness of such an ex ante injunction and give natural expression to my disappointment or **(p.235)** impatience at your not having been moved by the considerations that it would have made salient: at your not having responded to the expectations that my remark expresses.

Looking on your failure from a regulatory perspective, then, I cannot simply treat it as an occasion for resignation to brute fact or as an opportunity for indicating how you might have been a better agent in various ways. I can treat it only as an occasion for reminding you in an exhortatory spirit of what you could have done. Thus, I will naturally be led to use the only sorts of words available to me after the fact—words such as "You could have told the truth"—to give expression to the evocative or injunctive aspiration with which I might have said before the action: "You can tell the truth."

On this approach, my saying "You can tell the truth" prior to your acting does not merely register that, absent suitable exemption and excuse, you are disposed to tell the truth in many variations on the situation that continue to provide the same support for truth-telling. Uttered in the relevant context, "You can tell the truth" helps to elicit the very capacity it ascribes. The remark is not a descriptive utterance like "You have a ruddy complexion" that records a state of affairs that obtains independently of whether or not it is made. And nor is it a performative utterance like "I resign" that brings about the very state of affairs that it records (Lewis 1983b, Ch. 12). Rather, it falls somewhere in between the two. It is an evocative or exhortatory remark that helps to bring about—but does not guarantee to bring about—the state of affairs that it records (McGeer and Pettit 2015).

Saying "You could have told the truth" in wake of a lie amounts to treating you as an addressee of exhortation in the same way that anyone might have treated you prior to your offense. What the words convey is not, as in a non-naturalistic theory, that you had some special, non-natural ability that no causal factors could have withstood. What they communicate, rather, is that at the time you spoke you were a fit target for exhortation and injunction, unhindered by exempting or excusing factors, and that your actual failure to tell the truth does not negate that. The words convey something about you that depends on your **(p.236)** relationship to me and others and something, in particular, that allows me to remonstrate with you about your failure.[12]

Defending the theory against the alternatives

How does the theory supported here compare with alternatives? It scores in a general metaphysical fashion over any non-naturalistic approach, since it presupposes nothing that is inconsistent with the scientific image of the world. This is not going to be a knock-down consideration for those who are prepared

to countenance a non-natural, contra-causal sense of freewill. But it is a decisive objection from the point of view adopted in this work.

How does the naturalistic approach compare with the attitude-centered, naturalistic alternatives? There are three respects in which it ought to prove more appealing.

First, unlike those views, it preserves the idea, central to a long tradition, that in holding someone responsible for an offense we blame or censure the person, not for the attitude they displayed—or not just for that attitude—but for the deed they actually performed. We may condemn the careless driver who is about to drive through a red light for the attitude they take to traffic regulations, even if seeing a police car inhibits them. And we may equally condemn the careless driver who actually drives through, and perhaps injures a pedestrian, for their attitude. But we will also condemn the second driver for what they did. So, blaming that driver for that action cannot just consist in blaming them for their attitude.

Second, the current account gels more comfortably with the assumption, central to long religious tradition, that if you are fit to be held responsible for an offense, then you must have acted with full knowledge of the guilt and full consent of the will (Pettit 2007c). Those **(p.237)** conditions are intuitively satisfied if you stand in such a relationship to us that I can meaningfully credit you with a robust capacity to have registered and acted on the arguments against the offense and can meaningfully enjoin you to have performed better, treating you as a fit target for retrospective exhortation. But it is not clear that they will be satisfied just on the basis that you might have had better attitudes and might have been a better person.

But apart from these two, quite specific problems, there is a general question as to how far the attitude-centered approach scores above the current account in the very regard that may have seemed to make it more appealing. It supposes that there is nothing problematic about my claim that you could have had a different attitude. But that supposition itself raises a challenging query. Do I say that it was possible for you to have had a different attitude—say, for you to have been goodwilled—merely in the sense that this was not ruled out, and may even have been ruled in, by natural facts about your makeup and milieu? Or do I say this with an implicature of retrospective exhortation?

If I make just the first, non-exhortatory claim in holding you responsible for telling a lie, then holding you responsible in that way amounts merely to putting you among the goats rather than the sheep—distancing myself from you as from someone with whom a profitable relationship is unlikely (Scanlon 2008). But that really doesn’t answer to the phenomenology of blame, since it communicates distaste rather than impatience, frustration, and resentment (McGeer 2013). It

represents a radically revisionary account of what it is to hold someone culpable for what they did and to censure them for it.

Perhaps the intent of the approach, however, is that I make the second, exhortatory sort of claim in saying that you could have had a better attitude. Perhaps the idea is that in saying this I mean also to imply that retrospective exhortation, or some such attitude, is appropriate. T.M. Scanlon (1998, 22) seems to support such a reading when he says that because the attitude for whose absence we blame you is "judgment-sensitive"—dependent on your judgments about reasons—it is "up to you." And Angela Smith (2005, 263) supports a similar reading when she says that on the approach that she shares broadly with Scanlon, you **(p.238)** are "active, and responsible, for anything that falls within the scope of evaluative judgment." The fact that the attitude was up to you, or that you were active and responsible in endorsing it, would suggest that retrospective exhortation is indeed appropriate. But the mere fact that the attitude might have been different—or even that it is surprising that it wasn't different—does not suggest anything of the kind.

On the account presented here, it is likely to make perfectly good sense for those of us in Erewhon, or for those in the actual world, to blame one another for attitudes as well as actions. This will make sense if it is a matter of common belief that we are virtually pledged to follow certain epistemic standards—shared and routine standards—in how we form our beliefs, desires, or intentions, say, standards bearing on the import of certain data or desiderata. I may blame you for not living up to such a standard in the formation of your attitudes, say, for not being prepared to accept that a certain massacre occurred, or for not reacting with distaste to someone's cruelty. I may treat you as someone fit to be exhorted, retrospectively as well as prospectively, to respond attitudinally to compelling evidence of the massacre or to the unambiguous specter of cruelty.

But if the claim that you could have had a better attitude is supposed to be heard in this exhortatory way, then the grounds for adopting the attitude-centered theory disappear. If retrospective exhortation is allowed to give a richer sense to the claim that you could have had a better attitude, why shouldn't it be allowed to give a richer sense to the claim that in the actual behavior adopted, you could have done otherwise than you did? I conclude that as the approach defended here scores over non-naturalistic theories of responsibility, so, too, it has a clear advantage over naturalistic theories of an attitude-centered sort.[13]

(p.239) Three qualifications

In order to round out the account of responsibility supported by the narrative of this chapter, it may be useful to comment on three ways in which it is less committal than may initially appear. The first comment, which builds on some remarks earlier in the chapter, is that the capacity or disposition it postulates need not be as demanding as it seems. The second is that I may ascribe the

capacity to you and others, yet hold that you are not so deserving as others of being penalized for a failure to exercise it. And the third comment is that there is a sense in which I may hold you responsible even in the absence of the capacity normally postulated.

Taking up the first comment, the capacity or disposition I ascribe when I hold you responsible for telling the truth may require that you are likely to exercise it across relevant scenarios only at a relatively high level of dependability, not without exception. Again, the requirement is not that it is easy for you to tell the truth, only that you can get yourself to tell the truth: this, perhaps, by means of self-exhortation, and with the help of exhortation from others. Finally, the capacity that you are required to have need not be a durable disposition; it may depend for its appearance on contingent enabling conditions such as the level of exhortation to which you are exposed.

Just as you may have the capacity to live up to certain standards of desirability only in a relatively undemanding sense, so, to move to a second comment, your possession of that capacity may be consistent with not penalizing you as harshly as others for a failure to exercise it. While ascribing the capacity, and treating you as a conversable subject, I may recognize that there are factors that make its exercise and nurture particularly difficult. You may be lacking in opportunities to live up to those standards, say by avoiding crime, without losing out significantly in other respects; you may be exposed to very few role models who manage happily to live by those standards; or you may be embedded in a sub-culture where rival, in-group standards are given precedence.

Such problems would not constitute exempting conditions, by any account. Nor are they likely under any approach to count as excuses **(p.240)** that would let you off the hook for a failure of conformity; they will not count as regulation-resistant factors that would make exhortation pointless. But consistently with the theory endorsed, I may still think it important to take them into account in judging whether it is reasonable to impose extra sanctions of norm or law, whether on consequentialist, contractualist, or other grounds.[14]

The third comment on the line taken here is that I may often go through the motions of holding you responsible, when as a matter of fact I do not think that you currently have the capacity to live up to the standards I invoke. I may do this, because of a plausible belief that the best way to make you fully responsible in the longer run—think, for example, of the adolescent child—is to treat you as if you were responsible. The belief guiding my attitude is not that you are responsible in the sense of being fully fit to be held responsible but, in a criminological phrase (Garland 2001), that you are “responsibilizable.” You are capable of being made fit to be held responsible by being held responsible, so

that holding you responsible has a sensible developmental rationale (Pettit 2007c).

Notes:

(1.) They may not carry over smoothly, since doing good often makes more robust demands than doing bad (Pettit 2015a, c). This asymmetry between good and bad generates a similar asymmetry between holding someone responsible for good and holding them responsible for bad. The topic relates to the disposition-dependence of many of the goods you can bring about, which is discussed briefly in chapter 7, section 5. The complexity will not be pursued further here.

(2.) While different in the concrete version to be defended, the approach taken here broadly conforms to the abstract framework for conceiving of responsibility that Gideon Rosen (2015) describes as the alethic conception as distinct from the fittingness or moral conception. It denies that the relation in virtue of which it is appropriate to hold you responsible is primitive or sui generis, as in the fittingness view, or that it is a matter of the fairness of holding you responsible, as in the moral. It holds that what makes it appropriate to hold you responsible is just the truth of the thoughts or messages that adopting such a stance conveys.

(3.) For a discussion of how to think about variation on this front, see Pettit (2015c, Appendix II).

(4.) On the idea of being able to get yourself to do something, see Estlund (2014) and Southwood (2017).

(5.) In this case, I ascribe a disposition to tell the truth dependably or robustly but a disposition that may fail to be durable—that is, fail to be robust in the independent sense of surviving robustly over the ravages of time and other disruptive factors that do not count intuitively as distractors or disturbers. See Pettit (2015c); McGeer and Pettit (2017).

(6.) Victoria McGeer and I (2015) take the problem of explaining why “You could have done otherwise” does not just amount to saying it was a fluke that you did not do otherwise to constitute “the hard problem of responsibility.” The line taken in the text broadly follows the approach that we adopt in the paper of that name.

(7.) I may blame you unintentionally, if not unconsciously, in the sense in which blame involves just an attitude but not necessarily an action, not even the internal or mental act of assuming or taking up an attitude. I use the word “censure” here to refer essentially to the action of holding responsible; and this, in either an external or internal sense.

(8.) A comparable claim applies also in the case of avowing, but this will not be of concern here.

(9.) The idea of an ecological capacity is borrowed from Vargas (2013). The ecological idea ought also to appeal to Pamela Hieronymi (2007, 111) insofar as she takes the following to be true of moral capacities as well as capacities of other sorts: “Typically, our capacities develop as demands are put upon us to exercise them well—beyond our current ability,” where “the demands one is under remain insensitive to one’s own particular shortcomings; one’s capacities develop as one tries to meet them.”

(10.) The line taken makes good sense of the reactions we might have even in Frankfurt (1969) cases of the kind mentioned in Chapter 2 and earlier in this chapter. Suppose I know, but you don’t, that there is a back-up agency in place to ensure that you will tell the truth, even if you decide to lie or to mislead me otherwise. Still, I may exhort you to tell the truth, wanting you to do so spontaneously, and not as a result of that agency. Or suppose I know, and you don’t, that there is an agency in place to ensure that you will mislead me, even if you decide to tell the truth. Still, I may exhort you to tell the truth, wanting you at least to try to tell the truth: wanting you to mislead me, not willingly, but only as a result of the intervention of that agency. If I hesitate in either case to exhort you to tell that truth, that will only be because the exhortation would communicate a false message: that there is no one around to force your hand. But if I do exhort you in that way ex ante, saying “You ought to tell the truth”, then I will be able to say something similar ex post, communicating in the first case, that you ought to have told the truth willingly; in the second, that you ought at least to have tried to tell the truth.

(11.) I am indebted to Victoria McGeer for introducing me to this way of thinking about the regulative significance of excusing or exempting causes on the one side, non-exempting and non-excusing causes on the other. On this issue, and on the treatment of responsibility in general, I stick closely to our joint work in McGeer and Pettit (2015).

(12.) The approach counts, in a term from Bernard Williams (1995), as a “proleptic” theory insofar as it makes sense of blaming someone on the basis of the general effect of blaming in making a person responsive to others; faithful to the etymology of the word, it broadly involves treating something anticipated as already attained.

(13.) The action-centered approach also scores over the attitude-centered approach in other ways. It makes better sense of the idea that you may blame yourself for a failure, expressing an exhortatory attitude toward yourself. And it makes good sense of the fact that we blame corporate bodies for offenses committed in their name, recognizing that there is ground for exhorting them

prospectively and retrospectively to a better level of performance (Pettit 2007c; List and Pettit 2011).

(14.) The line in this paragraph reflects lessons that I have learned from Benjamin Ewing (2016). That line is particularly appealing on the assumption made throughout the text, that exempting and excusing factors do not come in degrees.