

From Moral Blame to Moral Wrongness¹

[I must apologize for the fact that this paper is very late, very rough, incomplete, and nearly devoid of citations. As a result, I recognize that I am deserving of the very attitude of moral blame that is the central theme of this paper. An unintended benefit of this situation is that the reader is likely to have this attitude in mind while reading the paper, and will thus be in a better position to evaluate the paper's central claims.]

The aim of this paper is to explore, in preliminary manner, an approach to understanding moral wrongness. The aim of this approach is to understand moral wrongness in terms of moral blame. It is widely agreed that there is a close connection between wrongness and blameworthiness. Sometimes the connection is understood as follows:

- (1) If an agent S performs an action A, then S is blameworthy for S just in case S's doing A is wrong and S lacks an excuse for doing A.

And sometimes this connection is understood as follows:

- (2) If an agent S performs an action A, then S is blameworthy for S just in case S's doing A is wrong and S is morally responsible for doing A.

Now if a connection of this kind exists, then it seems it could be used to shed light on one or other of the relata of the connection. Thus, if we begin with an independent understanding of moral wrongness and of excuses, then (1) will provide an account of blameworthiness. Similarly, if we begin with an independent understanding of moral wrongness and of moral responsibility, then (2) will provide an account of blameworthiness. Alternatively, if we begin with an independent account of moral wrongness and of blame, then (1) will provide an account of excuses, and (2) will provide an account of moral responsibility.

These kinds of strategy have been pursued in some detail. Here, however, I will be exploring a different approach. I will be exploring the project of understanding moral wrongness in terms of the other concepts—that is, of understanding wrong actions as the kinds of action which, if done responsibly or without an excuse, would be blameworthy. The goal of this approach is not simply to provide a *metaethical* account of what is involved in *saying* or *judging* an action to be morally wrong, but a *normative* account of the conditions under which an action *is wrong*. While I will not attempt to complete this task within the confines of this paper—doing

¹ I am very much indebted to Chandra Sripada, whose ideas greatly influenced my thinking in writing this paper.

so would require nothing short of a monograph—I will be offering some suggestions concerning how it might be carried out.

The first step in this project is to provide an account of *blame*. After considering two accounts of blame that are popular among philosophers in sections 1 and 2, and I will propose an alternative account in section 3. In section 4, I will raise and respond to a number of objections to my account. The next step will be to provide an account of *blameworthiness*. To do so, I will begin, in section 5, with a general account of what makes an attitude fitting. Then, in section 6, I will offer a general account of what makes an object worthy of a given attitude—as we will see, the question of fittingness and the question of worthiness are distinct. In section 7, I will combine my account of blame with my account of what makes an object worthy of an attitude to arrive at an account of blameworthiness. And in the final section, I will suggest how this account of blameworthiness could be used as the basis of an account of moral wrongness.

1. The Impaired Relationship Account

In *Moral Dimensions*, Tim Scanlon proposes the following:

To claim that a person is *blameworthy* for an action is to claim that the action shows something about the agent's attitude toward others that impairs the relations that others can have with him or her. To *blame* a person is to judge him or her to be blameworthy and to take your relationship with him or her to be modified in a way that this judgment of impaired relations holds to be appropriate (p. 128).

Later, Scanlon clarifies what it means for one's relations with a person to be modified "in a way that this judgment of impaired relations holds to be appropriate." It is, he explains, "to have attitudes toward him that differ, in ways that reflect this impairment, from the attitudes required by the relationship one would otherwise have with this person."

There is an immediate problem with this account. Not all ways of showing relationship-impairing attitudes are blameworthy. As an illustration, suppose Hannibal is a psychopath, and that initially his neighbor Louise is unaware of this fact. One day, however, she notices that he checks himself into the local Psychopath Treatment Clinic, where psychopaths are taught to control their violent impulses. Since only psychopaths check themselves into the Psychopath Treatment Clinic, Hannibal's doing so shows that he is a psychopath, and so it shows something about his attitudes that impairs the relations that others can have with him. If Louise recognizes that Hannibal's checking himself into the Clinic has this feature, then it follows from Scanlon's account that she thereby judges Hannibal to be *blameworthy* for checking himself into the clinic. And if, in response to this realization, she takes her relationship with Hannibal to be modified accordingly (e.g., if he comes to trust Hannibal less), then it follows from Scanlon's account that she blames him for checking into the Clinic. But these implications are highly implausible.

We could solve this problem by stipulating that the action blamed needs to show the relationship-impairing attitude *in the right kind of way*. It's not enough that the action constitute evidence, even conclusive evidence, for the presence of the relationship-impairing attitudes. Rather, the action must be *motivated* by these attitudes. Since Hannibal's checking into the clinic isn't motivated by his psychopathic attitudes, this revised account would avoid the implication that this action is blameworthy.

Even if we make this revision, however, serious problems remain. Suppose we take an agent's actions to show something about her attitudes that impairs her relations with persons *other than ourselves*. As an illustration, Consider Fred, who hates Canadians with a passion. He is the sworn enemy not only of all Canadians, but of everyone who is on good terms with Canadians. As a result, he is a very lonely and isolated individual, since he regards most people as his enemies. One day, however, he comes across a fellow Canadian-hater, Bill, whom he sees spitting on Canadians and burning Canadian flags. In this case, Bill's spitting on Canadians not only shows but is *motivated by* his hatred of Canadians. And this hatred impairs the relations that Bill can have with others—in particular, it prevents him from being on good terms with Canadians. Hence, if Fred realizes that Bill's spitting on Canadians has this feature, then it follows from Scanlon's account (even in the revised form just suggested) that Fred thereby judges Bill to be blameworthy for spitting on Canadians.

Now suppose, further, that, in response to the realization that Bill's relations with Canadians are impaired, Fred takes his relation with Bill to be modified in the following way: he ceases to regard Bill as an enemy, and so he ceases to have the hostile attitudes toward Bill that are required by the relationship of enmity that he would otherwise have with Bill. In this case, it follows from Scanlon's account that Fred thereby blames Bill for spitting on Canadians. And this implication is hardly plausible.

One might try to solve this problem by stipulating that, in order for person A to count as blaming person B, A must see B's actions as showing that B has attitudes that impair B's relations *not just with anyone* but, minimally, *with A*. But the resulting view still has problems. Consider Mary and Jane who are best friends. Their friendship revolves around two activities: playing the video game Grand Theft Auto (which they do every day at Mary's house) and observing meteor showers (which they are able to do only on rare occasion). Suppose, however, that recently Jane has ceased to enjoy playing Grand Theft Auto. But since she knows that Mary could never enjoy any other game nearly as much as Grand Theft Auto, Jane doesn't tell Mary about her changed preferences, and continues playing Grand Theft Auto with Mary. However, Mary begins to suspect that Jane may have ceased enjoying Grand Theft Auto, and so she decides to spy on Jane while Jane is playing video games privately at her own home. And she observes that Jane routinely chooses any other video game—even her least favorite video games—over Grand Theft Auto. When Mary sees that she and Jane have such divergent tastes in video games, she concludes that it no longer makes sense for them to play video games together.

Nonetheless, she wants to be friends with Jane, and to meet on occasion to observe meteor showers.

In this case, Jane's choosing every other game in preference to Grand Theft Auto shows, and is motivated by, a dislike of Grand Theft Auto that impairs her relationship with Mary — since it means that there is no longer any shared activity that they can engage in regularly and that they both enjoy. Hence, if Mary recognizes this, then it follows from Scanlon's account that she thereby judges Mary's private video game choices to be blameworthy. And if, in response to this realization, Mary takes her relationship with Jane to be modified accordingly—e.g., if she ceases to regard Jane as a suitable video-game playing companion—then it follows from Scanlon's account that Mary thereby blames Jane for her private video game choices. And this hardly seems plausible.

Perhaps we can do better if we turn from Scanlon's general account of blame to his account of *moral blame*. According to Scanlon, there are different kinds of blame corresponding to different kinds of relationship that can be impaired. Thus, there is a kind of blame that friends can have to other friends in response to impairments in their friendship, a kind of blame that colleagues can have to one another in response to impairments in their collegial relations, and similarly for other kinds of special relationship between persons. But there is a kind of blame that any person can have toward any other person, in response to impairments in a relationship that holds between any two persons, namely what Scanlon calls the *default moral relationship*. On Scanlon's view, to judge that someone is *morally blameworthy* for an action is to judge that this action shows that they have attitudes toward others that deviate from the default moral relationship. And to *morally blame* them for their action is to regard their action as blameworthy and consequently to hold attitudes toward them that differ, in ways that reflect this impairment, from the default moral relationship (p. 141).

What, then, is the default moral relationship? While other relationships, such as friendship, are constituted by the particular attitudes that the members of the relationship have toward one another, and hence only exist between persons having these attitudes, the default moral relationship is simply a “normative ideal . . . that specifies the attitudes and expectations that we *should* have regarding one another” (p. 139, emphasis added). It is important for Scanlon that the default moral relationship not simply involve the kinds of attitudes that are *morally required*, such as the standing intention not to harm others, to keep one's promises, to avoid lies, etc. The reason the default moral relationship can't simply consist in these required attitudes is that, on Scanlon's view of moral blame, blaming someone involves holding attitudes toward them that differ from those required by the default moral relationship. However, Scanlon insists that blaming someone should not involve any change in our conception of, or in our commitment to, our basic moral obligations toward them: even when we hold that someone's conduct is blameworthy, and we blame them accordingly, we should still regard them as having the same rights and we should still respect those rights (p. 142). And so we should retain the standing intention not to harm them, lie to them, etc. And the departure from the default moral

relationship can't involve a departure from these morally required attitudes. And so it follows that, in addition to these morally obligatory attitudes, there must be a further dimension to the default moral relation that is not morally required, and which we can suspend in relation to those whom we blame. This dimension, according to Scanlon, consists in a readiness to enter "specific relations that involve trust and reliance," such as agreements, joint projects, and friendships, as well as a disposition to help others with their projects when this can be done at little cost, to wish them well, and to take pleasure in their successes (p. 143-44). Thus, when we blame someone, while we should retain our standing intention not to harm them, cheat them or lie to them, we can suspend our readiness to cooperate with them, help them out, or become friends with them.

But now there is a problem. For since there are two dimensions to the default moral relation (the morally required dimension and the dimension that is morally ideal but not morally required), it follows that there are two ways in which someone's attitudes might impair the default moral relationship. First, one might impair this relationship by having morally impermissible attitudes, such as the intention to harm, lie or cheat. Alternatively, one might impair this relation by having attitudes which, though morally permissible, fail to live up to the ideal of the default moral relationship. Consider, for example, Henry the unwavering hermit. Henry has built himself a hut on the top of a high mountain, and surrounded it with signs that read "do not disturb." On the rare occasion when someone overlooks his signs and comes to his door seeking a friend or partner in some joint enterprise, he replies "sorry, but I'm a hermit. Please find someone else." Similarly, if someone attempts to help him with any of his projects, he says "thanks, but I'm a hermit, so I'd much rather be left to my own devices." In this case, it seems Henry's attitudes are perfectly permissible, but since they prevent him from entering special relationships with others, they impair the default moral relationship as Scanlon conceives it. Moreover, Henry's saying "sorry, I'm a hermit" in response to invitations to enter special relationships shows that he has such moral-relationship-impairing attitudes, and is motivated by these attitudes. And so it follows from Scanlon's account that if someone were to recognize that Henry's utterances had this feature, they would thereby judge him to be blameworthy for making them. And, if, in response to this recognition, they were to modify their relationship with Henry, ceasing to see him as a suitable friend, or partner in joint enterprises, or recipient of assistance in his projects, then it follows from Scanlon's account that they would thereby blame him. And this implications seems highly implausible.

So far, I have argued that Scanlon's account fails to provide *sufficient* conditions for blame. For I have argued that one can see someone else's actions as displaying attitudes that impair the default moral relationship, and adjust one's attitudes toward him or her accordingly, without blaming the person in question. I will now argue that Scanlon's account likewise fails to provide necessary conditions for blame. For one can blame someone without seeing this person's actions as displaying attitudes that impair the default moral relationship, as Scanlon conceives this relationship. To see why this is so, let R be the default moral relationship, as Scanlon understands it. And consider the following scenario. On his deathbed, Linda's uncle asks her to

promise him that she will scatter his ashes next to those of his wife. Since Linda's making this promise would mean a great deal to her uncle, and since fulfilling this promise would take only 10 minutes of her time, we may plausibly assume that the attitudes required by the default moral relationship would motivate Linda to make the promise, and then to carry out the promised action. And suppose Linda has these very attitudes, and hence that she makes the promise and carries out the deed.

Now suppose, further, that this is all observed by Jeremy, who is a classical act utilitarian. Jeremy recognizes that Linda's attitudes are perfectly in line R. Suppose, however, that Jeremy does not endorse R. Being an act utilitarian, Jeremy thinks that what Linda ought to have done is to make the lying promise to scatter the ashes (to make her uncle happy), and then flush the ashes down the toilet (so as to have a few extra minutes in which to raise money for Oxfam). Suppose further that, immediately after scattering the ashes, Linda is injected with a Utility Maximizing Serum which transforms her motivations so that they align with act utilitarianism, and that Jeremy knows all this. And suppose, finally, that Jeremy knows that he is the only one who knows about Linda's transgressions. It seems that, in such a case, Jeremy might nonetheless *blame* Linda for these transgressions, though he may take himself to have no reason to express this blame, since doing so would have no deterrent effect.

But Scanlon's view implies otherwise. For it implies that Jeremy can only blame Linda if he regards her actions as displaying attitudes that are incompatible with the default moral relationship, R. And yet Jeremy recognizes that Linda's actions and attitudes are perfectly in line with R.

There is an obvious solution to this problem. We must simply substitute replace *the default moral relationship* with the relationship that the blamer *regards as the default moral relationship*, or as the ideal standard used as the basis for moral evaluation. The result is the following view:

Modified Scanlonian View of Moral Blame: to judge that someone is *morally blameworthy* for an action is to judge that this action shows that they have attitudes toward others that deviate from *what the blamer regards as default moral relationship*. And to *morally blame* them for their action is to regard their action as blameworthy and consequently to hold attitudes toward them that differ, in ways that reflect this impairment, from what the blamer regards as the default moral relationship

Unfortunately, this modified view won't solve the problem. For, on this view, Jeremy will count as blaming Linda only if the following obtains:

- (a) Jeremy sees Linda's actions as displaying attitudes that deviate from those required by the kind of ideal moral relationship that *Jeremy* regards as the standard for moral evaluation.

(b) In response to this realization, Jeremy revises his own attitudes toward Linda so that Jeremy's attitudes themselves come to deviate from those required by this ideal relationship.

Unfortunately, however, these conditions are not satisfied in the case in question. For the ideal moral relationship that Jeremy regards as the standard of moral evaluation will be the kind of relationship in which utility-maximizers stand to one another. And, given the stipulations I have made about the case, act utilitarianism implies that, going forward, the attitudes and expectations that Jeremy and Linda should have toward one another are precisely that attitudes and expectations that utility-maximizers should have toward one another. Thus, act-utilitarianism does not imply that Linda's past transgressions should have any impact on the kind of relationship that they should have toward one another going forward. Thus, act-utilitarianism does not imply that Linda's past transgressions call for any revision, going forward, in their moral relationship to one another. And so it doesn't imply that Linda's actions or attitudes create an impairment in their relationship that calls for a response.

To sum up, Scanlon's view has serious difficulty in the case of blamers whose moral views are very different from those that Scanlon endorses. For such a blamer may blame an agent, even when the blamer recognizes that the agent's attitudes are perfectly in line with the default moral relationship, as Scanlon conceives it. We can solve this problem by moving to a revised version of Scanlon's view, according to which the relevant moral relationship is the standard that the blamer endorses. But this won't always work. For Scanlon's view requires a conception of the default moral relationship in which, when one person's actions display blameworthy attitudes, this always calls for others to respond by revising their attitudes so as to deviate from the default moral relationship. The problem, however, is that someone, such as an act utilitarian, might endorse a conception of the default moral relationship that does not have this structure.

2. The Reactive Attitude Account

I will now turn, briefly, to what is perhaps the most popular account of blame among philosophers, namely the reactive attitude account. In "Freedom and Resentment," Peter Strawson offers an account of an attitude that he calls *indignation* or, in its weaker form, *moral disapprobation*. Since then, many philosophers have suggested that blame can be identified with, or understood in terms of, Strawsonian indignation or moral disapprobation, or, alternatively, that Strawsonian disapprobation is conceptually tied to moral wrongness and responsibility in the manner in which blame is traditionally thought to be so tied. In this section, I will raise some worries about this kind of view.

Strawson understands indignation and moral disapprobation in terms of the more basic attitude of *resentment*. All of these attitudes belong to the category of *reactive attitudes*: they are reactions to the level of goodwill or its opposite that we perceive in others. They differ, however, in that resentment is a *personal* reactive attitude, whereas indignation and

disapprobation are *impersonal* reactive attitudes. To say that resentment is a *personal* reactive attitude is to say that it is a reaction to someone's level of goodwill toward ourselves. By contrast, to say that indignation and disapprobation are *impersonal* reactive attitudes is to say that they are reactions to someone's level of goodwill toward *some person or other*, or toward *some member of the moral community*. "Thus," Strawson says, "one who experiences the vicarious analogue of resentment is said to be indignant or disapproving, or morally indignant or disapproving. What we have here is, as it were, resentment on behalf of another... and it is this impersonal or vicarious character of the attitude, added to its others, which entitle it to the qualification moral."

Thus, resentment, indignation and disapprobation are all reactions to someone's lack of goodwill, or presence of ill will. But what kind of reactions are they? According to Strawson, they are themselves to be understood in terms of a withdrawal of goodwill. "Indignation, disapprobation, like resentment, tend to inhibit or at least limit our goodwill towards the object of these attitudes, tend to promote an at least partial and temporary withdrawal of goodwill; they do so in proportion as they are strong, and their strength is in general proportioned to what is felt to be the magnitude of the injury and to the degree to which the agent's will is identified with, or indifferent to, it." Thus, according to Strawson, to resent someone is to withdraw goodwill toward them in response to their lack of goodwill toward us, whereas to have an attitude of indignation or disapprobation toward someone is to withdraw goodwill toward them in response to their lack of goodwill toward some member of the moral community.

I believe that Strawson's account of moral disapprobation, like Scanlon's account of blame, fails both in providing necessary conditions and in providing sufficient conditions. The reason it fails to provide necessary conditions is this. In general, one can have a given attitude vicariously only if there is *someone* whom we take to be in a position to have that attitude in the ordinary, non-vicarious manner. Thus, if I'm watching you surfing, and I'm vicariously enjoying the activity, then I must think that you are in a position to *directly* enjoy the activity. Similarly, I can vicariously resent an activity only if I think there is someone who is in a position to resent it directly. However, I can have moral indignation or disapprobation for an activity which I believe that no one is in a position to resent. Suppose, for example, that I have been raised to believe that masturbation, or premarital sex, or marijuana smoking, are morally wrong. I then learn that a certain person, Nancy, indulges in all these activities. Given my moral beliefs, it seems I could very easily have an attitude of indignation or moral disapprobation toward Nancy for these activities. But I don't directly *resent* her for these activity—after all, I don't think these activities harm me in any way, nor do I take them to show a lack of concern for my rights or welfare. Nor is there anyone else whom I take to be harmed by her activities, or for whose rights or welfare her activities show a lack of concern. Hence, there is no one whom I take to be in a position to directly resent Nancy for her activities. And so it follows that I can't vicariously resent her for her activities. And so Strawson's account of moral disapprobation has the implausible implication that I can't morally disapprove of Nancy for her activities.

Some readers might not be moved by this example. For some might hold that only someone who was thoroughly confused could morally disapprove of someone for the kinds of harmless activities that figure in this case. However, there are other kinds of victimless wrongdoing of which one could clearly morally disapprove without confusion. Suppose I find myself in Parfit's non-identity case, and I must choose between Depletion and Conservation. And suppose I choose depletion. As a result, total welfare is significantly decreased, since those who live in the distant future fare much worse given Depletion than they would given Conservation. Nonetheless, since my choice between these outcomes affects who exists in the distant future, and there is no one who would exist in the distant future regardless of which of these outcomes I choose, there is no one who is adversely affected by my choice of Depletion. Hence, there is no one who is in a position to directly resent my choice of depletion. Consequently, no one who understands all the facts about the case could vicariously resent me for choosing Depletion. And so Strawson's account has the implausible implication that no one could have an attitude of moral disapprobation toward me for choosing Depletion.

Thus, it seems we can have resentment or moral disapprobation toward someone without feeling vicarious resentment. And so Strawson fails to provide necessary conditions for these attitudes. I will now argue that he likewise fails to provide sufficient conditions. For we can vicariously resent someone without feeling indignation or moral disapprobation.

To see why this is so, we should first note that it is perfectly possible to resent someone without thinking he has done anything wrong. For example, Jim might blame his classmate Lucy for failing to invite him to her party without holding that she is under any moral obligation to invite him to her party. But surely if we can *directly* resent someone without regarding them as having done anything wrong, we can also *vicariously* resent someone without regarding them as having done anything wrong. Jim's friend, for example, might vicariously resent Lucy for failing to invite Jim to her party without regarding Lucy as having done anything wrong. In such a case, it hardly seems that we should say that Jim has an attitude of indignation or moral disapprobation toward Lucy. And so it seems these attitudes can't be identified with vicarious resentment.

Could we identify blame with a special form of vicarious resentment, perhaps *vicarious resentment on behalf of everyone in the moral community*? I believe we cannot. To see why not, let's consider another non-identity case. Suppose Ronda must choose between the following outcomes:

Outcome A: There are exactly one million people, and each one spends eternity in hell.

Outcome B: The one million people from outcome A also exist in this outcome, but they spend eternity in heaven. There are an additional five million people, however, and each one spends eternity in hell.

Now suppose Ronda chooses outcome A, and that I am someone who is condemned to eternity in hell as a result of her choice. In this case, I might easily feel resentment toward Ronda. Of course, I would recognize that what she did wasn't wrong, and this awareness would be likely to temper my resentment. But I might still experience resentment toward her for condemning me to hell. Moreover, everyone else who exists is in the very same position I'm in, for everyone who exists was condemned to hell as a result of her choice, and would otherwise have enjoyed eternal bliss. And since everyone could resent the Ronda's choice, I could experience vicarious resentment from the point of view of anyone, and so I could experience vicarious resentment from the point of view of the entire moral community. And so even the modified version of Strawson's view implies that, in having this generalized vicarious resentment, I would thereby have an attitude of moral disapprobation toward Ronda. But this seems wrong. For it seems I could have this vicarious attitude without experiencing any moral disapprobation.

One might respond that I am neglecting an important facet of Strawson's account of indignation namely that of *demands*. Strawson says: "The personal reactive attitudes rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings toward ourselves... The generalized or vicarious analogues of the personal reactive attitudes rest on, and reflect, exactly the same expectation or demand in a generalized form; they rest on, or reflect, that is, the demand for the manifestation of a reasonable degree of goodwill or regard, on the part of others, not simply toward oneself, but toward all those on whose behalf moral indignation may be felt, i.e., as we now think, towards all men [sic.]" Perhaps, by appealing to this notion of demands, Strawson could avoid the implication that, in the case I just described, I would morally disapprove of Ronda for choosing outcome A. After all, it doesn't seem that I *demand* that she act otherwise.

Unfortunately, Strawson has such a thin notion of a demand that it can't do the required work. For he says "these attitudes of disapprobation and indignation are precisely the correlates of the demand in the case where the demand is felt to be disregarded. The making of the demand *is the proneness to such attitudes.*" But on this thin conception of a demand, it seems that, since I *was* prone to respond with vicarious resentment toward Ronda's for her choice, it simply follows that I demanded that she act otherwise. And so we can't avoid the implausible implication that I thereby disapprove of Ronda's choice.

In the next section, I will follow Strawson's lead in understanding blame in terms of demands, but, in contrast to Strawson, I will attempt to give an independent account of the relevant notion of a demand.

3: Toward and Adequate Account of Blame

Let's take seriously the idea that there is something like a demanding attitude that underlies the blame attitude, such that blaming someone involves seeing what they have done as inconsistent

with what we demand. To do so, let's consider what might be considered a primitive form of demand.

Consider the following case. Gaby the goose has a large brood of goslings. Reynard the fox would like to eat one of those goslings, while Gaby would prefer that he not do so. If they were to get into a fight, Gaby would most likely win, and she would prevent Reynard from eating any of her offspring, but both animals would be seriously injured in the process. If Gaby doesn't prevent him, he will eat only one of her goslings and be on his way. Given these facts, it would not seem rational for Gaby to fight Reynard. Indeed, even if she had only a single gosling, it probably wouldn't be rational for her to fight. For she can always have more offspring later, and the injury she would be likely to incur from a fight with the fox would result in a loss of reproductive fitness that would outweigh the loss of any one gosling. Hence, it seems that it will never, or almost never, be rational for Gaby to fight off foxes that want to eat her young. But if she consistently acts in this rational manner, she'll have a hard time passing on her genes.

Suppose however that, when Gaby is confronted with a fox, she enters a motivational state in which she is disposed to throw rational calculation to the wind and fight any potential threats to her young. In this case, it would not be in Reynard's interest to attempt to eat Gaby's young, for he would most likely fail in doing so and end up with an injury that would make him less able to acquire food from other sources. However, so long as Reynard is unaware of Gaby's motivational state, and regards her as a rational agent, he will assume that she will not fight him, and so it will be rational for Reynard to try to kill her young. Having this aggressive disposition, without revealing it to Reynard, would lead to the worst possible outcome for Gaby. For, since Reynard would be ignorant of her disposition, he would attack, and, because of her motivational state, she would fight Reynard and be injured in the process, thereby significantly lowering her reproductive fitness.

Thus, Gaby's disposition to fight Reynard will only be effective in preventing Reynard from attacking if Reynard is aware of this disposition. Thus, Gaby needs to make her motivational state manifest. She might do this, for example, by hissing at Reynard. Now suppose she does so, and that this hissing effectively communicates to Reynard that she will fight him if he attacks her goslings. In this case, in hissing at Reynard, she would be performing a communicative act the function of which is to prevent Reynard from attacking by informing him that his attacking would have dire consequences, and thereby giving him a compelling reason not to attack. Thus, Gaby's hiss will have many of the structural features of a demand. Compare it, for example, to a case in which a drill sergeant tells a private to do pushups. In so doing, the sergeant is performing a communicative act whose function is to get the private to do the pushups by letting him know that, if he fails to do the pushups, he'll be in trouble, and thereby giving him a compelling reason to do the pushups. Hence, she will be demanding that the private do the pushups. In much the same way, in hissing at the Reynard, Gaby can be seen as issuing a primitive demand that Reynard not attack her goslings.

Let's now think about what would be required in order for this hissing to successfully serve as such a demand. To do so, it must successfully indicate Gaby's motivational state: it must indicate to the Reynard that Gaby is in a kind of motivational state that would result in her acting against rational calculation if provoked. And, in order for Gaby's hiss to indicate this, it can't itself be simply the result of rational calculation. It can't be, for example, Gaby calmly and rationally reasons in a manner that can be represented by the following practical syllogism:

Desire: To indicate to Reynard that I am motivated to fight him if he attacks my goslings.

Belief: That hissing at Reynard would be an effective means to indicate to Reynard that I am so motivated.

Intention: To hiss at Raynard.

For suppose Gaby were *not* motivated to fight Reynard. Nonetheless, she would be no less likely to have the beliefs and desires that figure as inputs in this practical syllogism. And so she would be no less likely to engage in this kind of reasoning, and hence to intentionally hiss at Raynard as a result of such deliberation. But if being motivated to fight makes no difference to how likely one would be to intentionally hiss as a result of rational deliberation, then the mere fact that one intentionally hisses as a result of rational deliberation cannot serve as an indicator that one is motivated to fight.

Thus, in order for Gaby's hissing to indicate to Reynard that she is in the kind of non-rational motivational state that would result in attacking Reynard if provoked, it must itself be *produced* by this same kind of motivational state (or at least by a motivational state that is causally tied in the right kind of way to the motivational state that disposes her to attack if provoked). Thus, her hissing must be not a rational choice, but rather an emotional expression which is a symptom of an underlying motivational state. Hence, there must be some underlying motivational state (or a pair of causally connected motivational states) which disposes Gaby both to hiss at Reynard and to attack when provoked. Such a motivational or emotional state we may be called a *demanding mental state*. And insofar as Gaby's desire to fight Reynard, upon seeing him attack her gosling, is triggered by this demanding mental state, we may call it a *reactive attitude*.

On the picture I am painting, a mental state M counts as an attitude of *demanding* that individuals in class C act in manner phi just in case:

- (i) Those who are in state M are thereby motivated to act less favorably towards those in class C if they fail to act in manner phi.
- (ii) Those who are in state M are thereby motivated to act in ways that are (at least somewhat) reliable indicators of this very attitude.
- (iii) The function of state M is to prevent those in C from acting in manner phi.

And on this picture, a reactive attitude is the kind of mental state that is triggered by a demanding attitude when the individual who has this attitude learns that the demand has not been satisfied.

But while this may serve as an account of demanding attitudes and of reactive attitudes, it doesn't yet serve as an account of *moral* demands or of *moral* blame. There are, I believe, two things that are missing. The first is the *public* nature of the moral attitudes, and the second is connection between these attitudes and *moral motivation*. I will discuss these two elements in turn, and I will then argue that they are both aspects of a single structural feature of moral demands.

When Gaby demands that Reynard not attack her goslings, she is saying, in effect: "don't you dare attack them, or you'll get it from *me*." Hence, this is an entirely personal demand. But when we morally demand something, we are demanding it, as it were, on behalf of the community. Hence, we are not simply calling on the target of our demand to comply with it on pain of punishment; we are also calling on other members of the community to join us in making this demand.

This public character can be seen, perhaps not in the behavior of geese, but certainly in the behavior of chimpanzees. Consider the following example from de Wall:

Jimoh, the current alpha male of the Yerkes Field Station group, once detected a secret mating between Socko, an adolescent male, and one of Jimoh's favorite females. Socko and the female had wisely disappeared from view, but Jimoh had gone looking for them. Normally, the old male would merely chase off the culprit, but ... this time went full speed after the culprit and would not give up. [However,] before he could accomplish his aim, several females close to the scene began to "woaow" bark. This indignant sound is used in protest against aggressors and intruders. At first the callers looked around to see how the rest of the group was reacting; but when others joined in, particularly the top ranking female, the intensity of their calls increased until literally everyone's voice was part of a deafening chorus. The scattered beginning almost gave the impression that the group was taking a vote. Once the protest had swelled to a chorus, Jimo broke off his attack with a nervous grin on his face: he got the message. Had he failed to respond, there would no doubt be concerted female action to end the disturbance.²

Since male chimpanzees are larger and stronger than female chimpanzees, it would be pointless for a single female, on her own, to threaten Jimoh. But together, a group of females can easily overpower a single male, and so they are in a position to make demands of him.

² F. B. M De Waal, *Good Natured: The Origins of Right and Wrong in Humans and Other Animals* (Cambridge, MA: Harvard University Press: 1996).

Note, however, that in order for their threat to be effective, it must be credible to Jimoh. And this raises the same complication we saw in the case of Gaby and Reynard. Just as, under ordinary circumstances, it would not be rational for Gaby to carry out her threat and attack Reynard, it seems that, under ordinary circumstances, it would not be rational for any individual chimpanzee to attack Jimoh. After all, even if the group as a whole is sure to defeat Jimoh, any individual who participates in the attack incurs a risk of being injured in the process. Moreover, for any given individual, there is only a very low probability that her participation in the joint attack will make any difference to its success: if enough others attack, the attack will be successful without her. And if not enough others attack, then attacking would be foolhardy. Thus, it seems rational for each female chimpanzee to sit back and leave the attacking to others. And if all the chimpanzees reason in this way, then none will attack. And so if Jimoh regards the females as instrumentally rational, he won't be dissuaded by their threats.

Perhaps a precommitment mechanism could do the trick. Suppose the female chimpanzees were in an aggressive state of mind, which was manifested in their behavior, which would motivate them to attack Jimoh regardless of whether it's in their interest to do so. In this case, so long as Jimoh recognized that the females are in this state, he would find their threat credible.

But there remains a problem. While Gaby's precommitment mechanism is clearly adaptive, the precommitment mechanism we are now discussing would appear not to be. For suppose there were a single individual, Lisa, who lacked this mechanism, and who therefore had no disposition to attack Jimoh. In this case, Lisa would avoid the risk involved in attacking Jimoh, and this risk would be shared only among those who had the disposition to attack. Lisa would thus be a free rider, who would benefit from the behavior-controlling behavior of her fellow chimpanzees without taking the risks involved in carrying out collective punishments. And so it seems she would have an advantage in passing on her genes. And so it seems the kind of precommitment mechanism under consideration would not be evolutionarily stable.

We can solve this problem if we suppose that, when chimpanzees begin their calls of "woaow", they aren't simply inviting others to join the chorus; they are *demanding* that others join in the chorus. In other words, we can solve this problem if we assume that, in uttering "woaow," they are manifesting a state which disposes them not only to act less favorably toward *Jimoh* if he fails to call off his attacks, but also to act less favorably toward other chimpanzees if they fail to manifest this same state of demanding that Jimo call off his attack. That is, we can solve the problem if we suppose that the kind of demanding carried out by the chimpanzees has a kind of reflexive structure: that in demanding a certain kind of behavior, they are at the same time demanding that this very behavior be demanded. If there were a demanding state that had this structure, then we can see how it would be adaptive. For, while those who lacked this state might avoid the costs of engaging in the punitive behavior that this kind of state motivates, they would incur the costs of undergoing the punishments they would receive for failing to have this very state.

I suggest that one feature of moral demands is that they involve this kind of higher-order demand. However, I don't think this is enough to qualify a demand as moral. To see why not, imagine two rival gangs in a territorial dispute. When someone in Gang A steps into the contested area, someone in Gang B may demand that he leave, and may demand that others in gang B join in demanding that he leave (and hence that they be disposed to join in punishing him if he fails to leave). Even so, however, this demand is hardly a moral demand. So what's missing?

To identify the missing ingredient, let's first try to see what's deficient about the kind of state we have identified, with respect to how effectively it would serve the purposes of behavioral control. Suppose you have an attitude that motivates those who are in this state to punish individuals who fail to phi, and also to punish individuals who are not motivated to punish individuals who fail to phi. What kinds of behavior would the prevalence of such a state within a population motivate? For one thing, it would motivate the punishment of those who are caught failing to do phi. And for another, it would motivate individuals to phi *in situations in which they might be caught for failing to do so*. However, it would not provide individuals with any incentive to phi in contexts in which they could refrain without detection.

To succeed in motivating behaviors in such contexts, we need a mechanism that not only rewards phi-ing, but also rewards *being disposed to phi*. And this will be possible only if being disposed to phi is itself detectable. This will be possible if there were an attitude that disposes one to phi, and that also disposes one to act in ways that manifest this very disposition. Let us call such an attitude an attitude of *being committed to phi-ing*.

Now if the *demanding* phi motivated those who had this attitude to punish those who failed to be committed to phi-ing, then the prevalence of this attitude within a society would create an incentive to be committed to phi-ing, and this, in turn, would motivate people to phi even when failing to do so would go undetected. Such an attitude, therefore, could solve the problem of motivating undetectable acts.

This, I think, is the missing ingredient. The reason that the demand made by the member of gang A to the member of gang B is not a *moral* demand is that all that is being demanded is a kind of behavior. To satisfy this demand, it suffices that the rival gang member leave the contested territory simply in order to avoid a confrontation. He isn't expected to be non-instrumentally motivated to avoid setting foot in the contested territory. And so he is not expected to avoid such behavior in contexts in which there is no chance of his being caught. The gang members imply expect compliance, not proper motivation. And so their demand isn't moral.

I have argued that moral demands involve two important features: First, in demanding phi, we demand that others likewise *demand* phi. And second, in demanding phi, we demand that others be *committed* to phi-ing. We can unite these two features so long as we make the

plausible assumption that the same kind mental state which motivates one to punish others for failing to phi (and for failing to punish those who fail to phi), also motivates one to phi. Let us call an attitude that has both these features an attitude of *norm-internalization*. The attitude of internalizing the demand to phi can be defined by the following three features:

- (i) Those who have internalized the demand to phi are motivated to phi;
- (ii) Those who have internalized the demand to phi are motivated to punish those who have not internalized the demand to phi.
- (iii) Those who have internalized the demand to phi are motivated to act in ways that manifest this very attitude.

In this case, internalizing the demand to phi would, in virtue of features (ii) and (iii), constitute an attitude of demanding phi. And, in virtue of features (i) and (iii), it would constitute an attitude of being committed to phi-ing. Moreover, since internalizing the demand to phi involves demanding this very attitude, and since this attitude is both a demanding attitude and an internalizing attitude, it follows that internalizing the demand to phi would involve both a demand that phi be demanded and a demand that others be committed to phi-ing. Hence, this attitude would have both the features that I have argued are essential to moral demands.

On the account I am proposing, the fundamental attitude is that of norm-internalization, which is both a demanding attitude and a commitment attitude. Insofar as it is a demanding attitude, it involves a disposition to become motivated to punish those who violate the internalized norm. And when this disposition is triggered by learning that the norm has been violated, the resulting mental state is the attitude of moral blame. This mental state disposes those who are in it to punish the norm-violator.

This concludes my positive proposal. I will now consider three objections.

4: Objections and Replies

Objection 1: Your account doesn't jive with our experience of blame. While the goose and the chimpanzees in your examples may be motivated to attack those who contravene their demands, the attitude of blame experienced by human beings needn't involve any such aggressive tendencies. We can easily blame someone, such as a stranger who has committed a minor infraction, or a loved one who has committed a more serious infraction, without having any desire to harm them in any way.

Reply: All this is true, but it's completely compatible with the account I have proposed. All I mean by *punishing* someone for phi-ing is responding to their phi-ing by behaving toward them in a manner that adversely affects their interests, relative to the manner in which one would behave toward them if they had not phi-ed. Thus, physically attacking someone for phi-ing is one way to punish someone for phi-ing, but it is by no means the only way. Failing to do

something beneficial for the person that one would otherwise do also counts as punishing, in the relevant sense. Further, refraining from entering into beneficial relationships with that person, such as friendships or partnerships in joint projects, can be another way of punishing them. Further, telling others about the person's wrongdoing, or simply expressing a negative attitude about the person to others, in such a way as to make others less likely to enter such relations with the person, can be another way of harming the person. And these *are* things we are generally motivated to do when we blame someone. (Note, further, that it doesn't matter, from the point of view of my account, whether we *conceive* of these actions as punishments, or whether we *intend* to adversely affect those whom we blame.)

Objection 2: The model you propose assumes that there are states with various motivational features, such that being in these states is obvious to others. But it's unclear how any such state could arise in nature. After all, there could be an alternative *mimicking* state that appeared just the same, but that lacked the same motivational tendencies. Such a state would have all the advantages of the state being mimicked, but since it would not involve any disposition to act against rational self-interest (e.g., by complying with norms when failure to do so would go undetected), it would be more adaptive. Hence, we should expect that, over time, those with the mimicking state would come to dominate.

Reply: The model does not require total transparency. All that it requires is that individuals in the population have better-than-chance odds of identifying those with the motivations in question. Consider, for example, the original case involving Gaby and Reynard. Suppose there are two kinds of geese: geese that genuinely become aggressive when foxes threaten, and are disposed to fight any attacking fox, and that geese aren't genuinely aggressive but pretend to be. And suppose these two types of geese are equally common. Suppose, however, that Gaby is a goose of the first kind (i.e., she is genuinely aggressive), and Reynard has a better-than-chance ability detect genuinely aggressive geese. In this case, there will be a greater than 50% chance that Reynard will guess that Gaby is probably really aggressive. And (depending on the probabilities and the payoffs) this guess could easily make it rational for Reynard to back down. Similarly, when confronting geese that are not genuinely aggressive, Reynard will have greater than 50% chance of guessing that they are probably only pretending to be aggressive, and hence that they will probably not attack. And this guess could easily make it rational for Reynard to attack the goslings of the mimicking goose. And these differential probabilities of attack could easily give the genuinely aggressive geese a selective advantage over the mimickers. More generally, mimicry is seldom perfect, and so long as it isn't perfect, mimickers may be at a disadvantage.

Objection 3: Your model assumes that, among the motivational states that are detectable by others is the state of being motivated to follow certain kinds of norms. But if a disposition to follow norms is detectable, that fact alone could explain how moral motivation could be genuinely stable. For, in world in which people can detect morally motivated individuals, everyone will have prudential reason to selectively cooperate with those with such motivations.

Hence, those who are morally motivated will have more access to cooperation, which will make them more successful. Consequently, the other features in your model, such as dispositions to punish, will not be necessary to motivate moral behavior.

Reply: If the only kind of behavior that needed to be prevented were *defection* in scenarios that have the structure of the prisoner's dilemma, then the transparency of moral motivation would suffice. For if it were clear to others who the defectors are and who the cooperators are, then the cooperators would have self-interested reason to interact exclusively with other cooperators, and the defectors would be stuck interacting only with other defectors, and so the cooperators would have a major advantage over the defectors. However, there are lots of actions that others might want to prevent, but where others would have no prudential reason to avoid interacting with those who engage in such behaviors (at least apart from any fear of punishment). Suppose Betty likes stabbing people in the face with forks while they sleep (perhaps because, like the Ice Queen, she wants to be the fairest of them all). However, while others are awake, she behaves just like everyone else. In particular, she acts just like everyone else when engaging in joint activities. Consequently, even if others know that she is motivated to stab them in their sleep with forks, they will have no reason to regard her as a worse partner in joint activities than anyone else. And so they will have no direct prudential reason to choose someone else over her as a partner in such activities.

Hence, while there may be some moral norms (such as the norm prohibiting defection) such that being transparently motivated to follow these norm would by itself give others a prudential reason to want to interact with you, there are plenty of other important moral norms that don't have this feature. In order for individuals to be effectively motivated to follow these other norms, some other mechanism is required. And the mechanism of norm internalization that I have sketched could play this role. For such a mechanism could be evolutionarily stable, and it could motivate compliance with a much broader range of norms.

5: What Makes an Attitude Fitting?

In order to go from an account of blame to an account of blameworthiness (and hence to an account of moral wrongness), I need some account of what makes an attitude *fitting*. In this section I will briefly consider two such accounts that I reject, and then offer an alternative account.

5.1 The Cognitivist Account of Fittingness

One account is the *cognitivist* account. On the cognitivist account, emotions involve a cognitive component. This is normally understood as a belief or judgement, or some related state such as a *contrual*. On most cognitivist views, this cognitive state isn't all there is to an emotion. An emotion will also have other aspects, such as motivations or dispositions to act in various ways, and perhaps also certain subjective feelings. And these behavioral dispositions and feelings are caused by the cognitive state. Thus, fear might involve the belief that the object

has a certain property (such as the property of being dangerous) which causes an unpleasant feeling and a disposition to flee the object.

Cognitivist accounts of the nature of emotions typically go hand in hand with a cognitivist account of the fittingness conditions for emotions. Those who hold that having emotion E toward some object x involves a specific judgement, namely the judgement that object x has property P, typically maintain that this emotion is fitting just in case x really has property P. Thus, for example, those who maintain that fearing x involves the judgement that x is dangerous typically hold that it is fitting to fear x just in case x really is dangerous.

As D'Arms and Jacobson have pointed out, *sentimentalists*, who aim to provide an account of properties in terms of fitting attitudes, have reason to avoid such cognitivist accounts of fittingness. Consider blame. If blaming someone for an action involved some kind of judgement or construal, what would this judgement or construal be? The most natural answer is that it is the judgement or construal that the action is wrong. But if we give an account of blame in terms of the judgement or construal that an action is wrong, then we can't, without circularity, give an account of wrongness in terms of fitting blame.

This objection may not be fatal, even for the sentimentalist who aims to understand wrongness in terms of fitting blame. For perhaps blame does indeed involve some judgement, but that this is not a judgement of moral wrongness. Indeed, the accounts of blame proposed by Scanlon and Strawson that we considered above both have this feature. On Scanlon's view, blaming someone involves the judgement that she has acted in a way that displays attitudes that impair some kind of relationship. And on Strawson's view, blaming someone involved the judgement that their actions display a lack of goodwill. Thus, while both these accounts understand blame in terms of some kind of judgement, neither one understands it in terms of the judgement that the act in question is morally wrong. And so neither one would give rise to circularity when combined with a sentimentalist account of wrongness.

But there is another problem with the cognitivist view of fittingness. To illustrate this problem, let us suppose, for the sake of argument, that the cognitivist view of emotions is correct, and that fear essentially involves a particular judgement, namely the judgement that the object feared is dangerous to the blamer. But now let us suppose that there is a race of beings, call them *schmumans*, who have an emotion that is very much like human fear, but differs in one respect. Like fear, schmear is accompanied by similar affect, physiological changes, and by a disposition to flee the object of the emotion. However, the emotional mechanism belonging to schmumans is very ancient, and evolved long ago, at a time when their ancestors didn't have sophisticated concepts such as the concept of *dangerous*. Instead, it evolved when they had only much simpler recognitional concepts, such as the concept *snake*. As a result, the schmear mechanism is set up in such a way as to trigger the schmear reaction in response to the judgement that a snake is present.

Now suppose that, over time, one species of snakes, the schmarter snake, loses its venom and ceases to be dangerous. Is schmeiar fitting toward schmarter snakes. It's true that, in having schmeiar toward a schmarter snake, one is judging it to be a snake. But it is also true that, in having schmeiar toward a schmarter snake, one is disposed to act towards it in ways that only make sense on the supposition that it is dangerous. Hence, one is disposed to treat it as dangerous. And, in *this* respect, it seems one getting things wrong. It seems to me, therefore, that there is an important respect in which one's attitude of schmeiar toward the snake is not fitting.

One might think that is simply a case where the schmumans simply have pragmatic reason to not to experience schmeiar. And pragmatic reasons, one might maintain, are the wrong kind of reasons to bear on fittingness. But I think this is a mistake. The fact that someone will give me a million dollars if I refrain from having the attitude of schmeiar toward a schmarter snake is clearly a reason to avoid having this attitude that does not bare on the fittingness of this attitude. But what's wrong with having schmeiar toward a harmless object isn't just that having this attitude won't have good consequences (or that it will have bad consequence). Rather, the problem is that having this attitude toward schmarter snakes consists in part in having certain desires, such as the desire to flee the schmarter snake, *and these desires make no sense*.

Here's another way to put the point. Beliefs aren't the only attitudes that can be fitting. Desires, intentions, ect. can also be fitting. Hence, if an emotion involves not only beliefs but also such conative components, then it would seem that, in order for the emotion as a whole to be fitting, these conative components would need to be fitting as well.

Thus, even if emotions have a cognitive component, there is reason to doubt that this component can suffice to explain the fittingness of the emotion. Other components, such as the conative component, may be relevant as well. And if cognitivism is false, and emotions don't include a cognitive component, then perhaps the conative component may fully explain the fittingness of emotions. This is the view I will explore in the next section.

5.2 The Conative Account of Fittingness

If the fittingness conditions for attitudes can't be understood in terms of their cognitive component, perhaps it can be understood in terms the desires or behavioral dispositions they involve. Consider, once again, the case of fear. Fearing x involves, among other things, a disposition to flee x. So perhaps we can understand the fittingness of fear in terms of the appropriateness of the kinds of behaviors that it disposes one to do. Here's one version of this view:

Simple Conative Account: it is fitting for person S to have attitude A toward x just in case S has reason to act in the ways in which having attitude A toward x motivates one to act.

But there is an obvious problem with this view. Sometimes an attitude seems fitting, even when acting in the manner in which this attitude disposes one to act clearly would not be fitting. Suppose, for example, that I am confronted by man-eating tiger. Suppose I know that if I ignore the tiger he has a 50% chance of killing me, but that if I attempt to flee the tiger, or act in any of the other ways that are characteristic of fear, then it has a 100% of killing me. In this case, it seems that it would be perfectly fitting for me to fear the tiger, but that I do not have sufficient reason, or indeed any reason, to attempt to flee the tiger, or to do anything else that my fear of the tiger motivates me to do.

There are, however, more sophisticated versions of the conative account of fittingness that can solve this problem. Consider, for example, the following:

Sophisticated Conative Account: it is fitting for an organism S to have attitude A toward object x just in case x stands in some relation R to S such that, under ordinary circumstances, the fact that a given object stands in relation R to oneself provides sufficient reason to act in the ways in which one is typically motivated to act in virtue of having attitude A toward the object in question.

Moving to this sophisticated view solves the problem raised by the tiger case. For, while one may have no reason to flee the tiger, or to act in any of the ways that fear motivates one to act in, the tiger does stand in a relation to one (namely the *dangerous-to* relation), such that, under ordinary circumstances, the fact that an object stands in this relation to oneself provides a sufficient reason to act toward the object in the ways typically motivated by fear.

But there is a problem. Some emotions motivate us to act in ways that seem irrational. Sadness motivates us to sob, to lie in bed all day, and to act in a generally lethargic manner even when there is no one around to observe our behavior. And since there doesn't appear to be any good reason to act in these ways, and there is plenty of reason not to act in these ways, sadness seems to motivate us to act in ways which, even under ordinary conditions, there isn't sufficient reason to act in. And so the behavior based view seems to imply, counterintuitively, that sadness is never fitting.

The reader might wonder whether this point undermines the argument of the previous subsection. There I argued that shame would not be fitting toward harmless smarter snakes, since this attitude would involve a desire to act in ways that make no sense. But now I'm conceding that sadness can be fitting, even when it motivates one to act in ways that are irrational. There is, however, no inconsistency here. For actions that are not rational can nonetheless make sense. If we understood the function of sadness, we might see that the kinds of actions it motivates make sense in light of this role, even if they are not rational. Similarly, on the view of blame I have suggested, if we understand the role of blame (as an attitude that plays a role in motivating behavior), then the costly punishments this attitude motivates may make sense in light of this role, even if the punitive acts are not rational.

To spell this out in greater detail, I will need to offer an account of fittingness in terms of functions.

5.3: The Proper Function Account of Fittingness

Let's again consider fear. It seems that it is fitting for someone to fear a given object just in case this object is dangerous to this person. Thus, the relation *is dangerous to* is the relation in which an object must stand in to an organism in order for it to be fitting for this organism to fear this object. Hence, we may say that this relation is the *fit-making* relation for fear. More generally, for any attitude A, relation R is the fit-making relation for attitude A just in case, necessarily, (it is fitting for a person S to have attitude A toward an object x if and only if x stands in relation R to S).

My goal will be to provide a general account of what makes a given relation the fit-making relation for a given attitude. As a first step to providing such an account, I propose the following:

Proper Function Account of the Fit-making Relation: For any attitude A and relation R, R is the fit-making relation for R just in case it is the proper function of the mechanism that controls the production of attitude A to track relation R.

To say that the mechanism that produces an attitude *tracks* a given relation is to say that this mechanism disposes its possessor to have this attitude toward all and only those objects that stand in this relation to its possessor. Thus, to say that the mechanism that controls fear production tracks the *is-dangerous-to* relation is to say that this mechanism disposes one to fear all and only those objects that are dangerous to oneself.

But what does it mean for tracking a given relation to be the *proper function* of the mechanism that controls the production of a given attitude? Here's a fairly natural suggestion:

Simple View: For any attitude A and relation R, the proper function of the mechanism that controls attitude A is to track relation R just in case:

- (i) This mechanism tracks relation R.
- (ii) This fact explains why this mechanism is conducive to reproductive success, and so it explains the existence or maintenance of this mechanism via natural selection.

Unfortunately, there is a problem with the simple view. The problem can be seen in cases where there are two relations to organisms that happen to be coextensive in the environment of the organisms in question. Suppose, for example, that there were a species of organisms that happen to be vulnerable only to snakes, and so the only things that are dangerous to them are nearby snakes. Now let R1 be the relation *is-dangerous-to*, and let R2 be the relation *is-a-snake-that-is-near-to*. Now let us consider two different attitudes, A1 and A2, each of which we may suppose

to track both of these relations. Let A1 be a fear-like attitude. In particular, let A1 be an attitude that disposes those who have this attitude to flee from the objects of this attitude. And let A2 be an attitude with a very different behavior profile. Among the behavioral dispositions associated with A2 is the disposition to exclaim “behold, a nearby snake!” as well as the disposition to accept bets on the proposition that there is a nearby snake.

Suppose both these mechanisms track both R1 and R2. Now it seems the fact that the mechanism that produces the fear-like attitude A1 tracks R1 (the dangerous-to relation) could explain its existence or maintenance. And since R1 is coextensive with R2, it seems the fact that this mechanism tracks R2 could likewise play a role in explaining its existence or maintenance: this mechanism persists because it tracks nearby snakes, and these happen to be dangerous. Hence, the simple view will imply that it is the natural function of this mechanism to track not only the dangerous-to relation, but also the nearby snake relation. And to the view under consideration implies that that R1 and R2 are each fit-making relations for attitude A1. For similar reasons, this view likewise implies that R1 and R2 are each fit-making relations for attitude A2.

But this seems like the wrong result. It seems that what makes the fear-like attitude fitting is the *is dangerous to* relation, and that what makes A2 fitting is the *is a nearby snake* relation. For it seems that it’s fact that something is dangerous that makes it fitting to have an attitude that disposes one to flee from it, and it’s the fact that something is a nearby snake that makes it fitting to have an attitude that disposes one to call it a nearby snake and to bet on its being a nearby snake, etc. And so we want an account that will enable us get this result, and, more generally, to discriminate among the fit-making relations of attitudes that track contingently coextensive relation.

I propose the following:

Sophisticated View: For any attitude A and relation R, the proper function of the mechanism that controls attitude A is to track relation R just in case:

- (i) This mechanism tracks relation R.
- (ii) This fact explains why this mechanism is conducive to reproductive success, and so it explains the existence or maintenance of this mechanism via natural selection.
- (iii) There is no further relation, R’, such that this mechanism tracks R’, and such that it is only in virtue of the correlation between relations R and R’ that the mechanism’s tracking R explains why it is conducive to reproductive success.

By adopting this sophisticated view, we can avoid the conclusion that the mechanism controlling A1 tracks R2. For it is only in virtue of the correlation between R1 and R2 that A1’s tracking R2

explains its existence or maintenance. And, in a similar manner, we can avoid the conclusion that the mechanism that controls A2 tracks R1.

Some readers may reject the account I have sketched, because some readers will deny that questions about reproductive fitness can have any bearing on normative questions such as questions about the fittingness of attitudes. But readers who hold such a view can still adopt some form of proper function account. They simply need some other way of understanding proper functions. And this seems possible. For example, long before we knew anything about natural selection, we held that the function of the heart is to pump blood. For we held that pumping blood is what the heart *is good for*. In other words, we held that the fact that the heart pumps blood explains why it's good for someone to have a heart, or why having a heart is beneficial. One possibility is that our prior conception of proper functions is fundamentally at odds with the evolutionary perspective. Another possibility, however, is the evolutionary account, in terms of reproductive fitness, can be seen as a precisification of the pre-Darwinian conception. This could be true, for example, if both pre-Darwinians and the Darwinians agree that the function of the heart is *what it's good for*, or *the effect of the heart that explains why it's beneficial to have a heart*, and the Darwinians have simply provided a further interpretation of what it is for a trait to be beneficial—to be beneficial, on the Darwinian view, it to be conducive to reproductive success.

If this is right, then we can arrive at a more theory-neutral version of our account of proper functions by replacing “conducive to reproductive fitness” with “beneficial to the organism.” In so doing, we arrive at the following view:

For any attitude A and relation R, the proper function of the mechanism that controls attitude A is to track relation R just in case:

- (i) This mechanism tracks relation R.
- (ii) This fact explains why this mechanism is beneficial to its possessor.
- (iii) There is no further relation, R', such that this mechanism tracks R, and such that it is only in virtue of the correlation between relations R and R' that the mechanism's tracking R explains why it is beneficial to its possessor.

Thus, on this theory-neutral version of the view, we can say that it is the proper function of the mechanism that controls fear to track the *is-dangerous-to* relation so long as this mechanism in fact tracks this relation, this fact explains why this mechanism is beneficial to those who have it, and this benefit isn't explained by a correlation between this relation and some other relation.