The Alethic Conception of Culpability

Gideon Rosen
Princeton University

(Rough Draft for the Columbia Legal Theory Workshop and the Yale Seminar on Law and Philosophy. Notes and references incomplete. Please do not cite.)

January 10, 2006

This paper addresses the following question: What is it for a person to be morally blameworthy (or, as I shall sometimes say, culpable) for an action? As I understand it, this question calls for a definition, not of the English word, but of the moral relation itself. We want an account of *what it is* for a person to be culpable for his bad conduct. This demand for definition is in part purely theoretical. Culpability is clearly an important moral relation, and any fully worked-out moral theory will say what there is to be said about its nature. However, my interest in the definitional project is not simply theoretical.

The central problem in this part of moral philosophy is to provide an account of the *conditions of culpability*: the conditions under which a person is properly blamed for his bad actions and their consequences. We begin the inquiry with a range of views about this. We have opinions about who is responsible for what in particular cases. And we have a practice of holding people responsible which plausibly embodies an inchoate theory of the conditions of culpability. (Following J. L. Austin, it is often said that this theory is best conceived as a theory of the excuses — the conditions under which a person is *not* responsible for his bad behavior.) Now suppose we want to know

whether our everyday opinions on these matters are correct. Suppose we want to know whether some ordinary excuse really does excuse bad conduct. Or more interestingly, suppose we want to know whether some novel putative excuse should be reckoned genuine. (For philosophers, one salient example is the alleged excuse that everyone would always have if physical determinism were true.) Suppose it is not immediately obvious whether this alleged excuse is a good excuse. How should we approach the question? We might simply eyeball the proposal in the hope that some clear intuition will emerge. But it would be disappointing if this were the best that we could do. I am interested in the analysis of blameworthiness because I want to know how we might *argue* for the cogency of a novel excuse. There is of course no guarantee that a suitable definition will resolve every hard question of this sort that might arise. But it would be an excellent thing if we could gain some rational leverage on such questions, and as we shall see, there is reason to hope that a definitional inquiry may provide some guidance. That, at any rate, is the thought that guides the present exercise.

1. Preliminaries.

If word structure is any guide, we may begin with the following working hypothesis. For a person to be blameworthy for an act (or more generally, for an occurrence) is for a certain negative response to the act[1] — call it "blame" — to enjoy a certain positive normative status — call it "appropriateness". Read nothing at all into the latter word at this point. It is simply a label for the normative status that blame has when and only when it is deployed in response to a genuinely blameworthy act. A complete definition of blameworthiness will then have two components: an account of *blame* and an account of what it is for blame to be *appropriate*.

Now it is a commonplace that one way to blame a person for an act is resent him or to feel indignant towards him for having done it. (If the act is yours, then you blame yourself in this way by feeling guilty.) In what follows, I make a stronger assumption. Following Strawson[2], Gibbard[3], Wallace[4] and others, I regard these so-called *reactive attitudes* or emotions as the primary forms of blame, and this in two senses. I suppose first that the other forms of blame are to be explained in terms of the reactive attitudes. I may blame Woodrow Wilson for the disastrous terms of the Treaty of Versailles without *feeling* much of anything. But it is plausible, as Wallace suggests, that this "cool blame" simply amounts to the judgment that indignation would be appropriate responses to Wilson's actions. Second, and more importantly, I suppose that all questions

[1] Strictly speaking, the response is not simply a response to the act. It is an attitude towards the agent and the act taken together. Blame is a three-place relation: X blames Y *for Z*. In this respect blame differs from attitudes that focus strictly no the *act* (hate the sin) and from those that focus strictly on the agent (hate the sinner). Many confusions in the theory of responsibility derive from the mistaken assimilation of blame to one or another of these agent-directed or act-directed sentiments. (Say more.)
[2] "Freedom and Resentment," in *Freedom and Resentment and Other Essays* (ref.)
[3] *Wise Choices, Apt Feelings,* Harvard University Press (Cambridge, MA, 1990).
[4] *Responsibility and the Moral Sentiments,* Harvard University Press (Cambridge, MA, 1994).

about the appropriateness of blame (and hence about blameworthiness) reduce to questions about the appropriateness of the reactive attitudes. If it is appropriate to blame Wilson for insisting on ruinous terms at Versailles, then this is so in virtue of the fact that indignation would be an appropriate response to his activity. These writers suppose, in other words, that to be blameworthy just is to be "resentment worthy", and I shall endorse this assumption in what follows.

Given these assumptions, a complete account of these matters will require (a) an account of the reactive attitudes (of which resentment will serve as the paradigm), and (b) an account of what it is for these attitudes to be appropriate. The main business of this paper is to explore a promising approach to these questions, an approach I call the Alethic Conception of Culpability.

2. Gibbard's Proposal.

The best way to see the point of the Alethic Conception is to consider some of the alternatives to it. So let's begin with a straightforward proposal due to Allan Gibbard.

According to Gibbard, an act is blameworthy just in case it would be *rational* for the agent to feel guilty and for others to resent him for having done it. Gibbard goes on to give detailed accounts of guilt and resentment. But for our purposes the important proposal is that the relation we have called "appropriateness" is to be identified with a certain species of rationality.

Which species? Gibbard cannot define the notion that interests him; it is too fundamental for that. Instead he offers the following informal gloss:

> Other phrases may capture this notion better. I have freely substituted talk about what it "makes sense" to do, to think, and to feel about things. Alternatively we might talk of what one "ought to do, think or feel" and

> explain that the ought is not a moral one. With feelings and beliefs we can talk of what states of mind are "warranted" or "well-grounded" or "apt"; with acts we can talk of the "best thing to do". We might simply talk of "the thing to do" or "the thing to feel" about something. If a flavorless recommendation on balance can be found in any of these terms, then that is what "rational" shall mean in this book.[5]

Now in fact there are important differences among these idioms. In particular, it is one thing to call an emotion "apt" or "well-grounded", something else to say that one ought to feel it, or that it is "the thing to feel". Idioms of the latter sort dominate Gibbard's lists. As he notes, they are terms of *on balance* or *all things considered* recommendation. And as Gibbard also notes, this suggests a problem for his account.

Cleopatra is vicious but perceptive. She responds brutally to perceived disloyalty among her courtiers, and when there is disloyalty she is likely to sense it. One day, on a whim, she orders the execution of a peasant for no good reason. We may stipulate that she has no excuse, and hence that she is blameworthy. Question: Is it rational in Gibbard's sense for her courtiers to respond with indignation? Gibbard says, "yes", as he must:

> Perhaps we can call anger irrational for the courtier, just because of the bad consequences he knows it must bring. Perhaps we can say that "it makes no sense" for him to be angry, because anger would be disastrous. That, however, is a very different judgment from the one I want to pursue. … In the relevant sense it would be rational for the courtier to be angry in that his anger is well founded or warranted, because Cleopatra has acted outrageously without excuse. (Ref.)

[5] *Wise Choices*, ref.

Gibbard's position, then, is that while the consequences of the courtier's response may be relevant to some forms of rational assessment, they do not bear on whether the response is rational in his intended sense.

Now it is clear that any normative relation suitable for the analysis of blameworthiness must have this feature. Whether Cleo is culpable for her transgression depends entirely on facts that were in place when she acted, and not at all on the downstream consequences that may attach to our responses. The trouble is that Gibbard's word "rational" and the synonyms he supplies for it are nearly all, as we have noted, terms of *on balance* appraisal. Consider the linguistic strain that shows in Gibbard's preferred response to the Cleopatra problem.

> In the case of the courtier and the queen, even though it is rational for him *to be angry* with her for ordering the execution unjustly, it may also be rational for him to *want* to ingratiate himself with her, for his own good and that of others. If anger would prevent that, then it may be rational for him to *want* not to be angry with her. In such a case, it is rational to be angry, but also rational to *want not to be angry*.

We need not deny that it can be rational to want to do something that it would not be rational to do, as when wanting will bring good consequences but doing will court disaster. But here it sounds bizarre to say that while it is rational for the courtier to want not to be angry, anger is nonetheless, *in the same sense*, the rational thing for him to feel. Suppose he struggles to stifle his incipient emotion. What justifies the effort? Not the thought that it is rational for him to *want* not to be angry. He has already achieved that condition. He already *wants* not to be angry. No; surely it makes sense for him to wrestle with his anger precisely because, in the circumstances, anger is *not* the thing to feel.

Gibbard might concede the point. He might concede that in his book, a range of idioms that normally connote an all things considered practical evaluation are being used in a special "some things considered" sense. In this special sense it can be rational to do something even when the answer to the practical question "Should I do it?" is clearly *No*. The challenge is then to say something illuminating about this special sort of rationality, and in particular about the range of considerations that determine whether a contemplated act or attitude would be rational in the special sense.

The Cleopatra problem is an instance of a general phenomenon. Suppose you have excellent evidence for some proposition *p*, but that believing *p* would kill you. Then in one sense it would be rational for you to believe that p, but in another sense you have decisive reason not to believe it. Suppose that some prospect *S* would be better for you than any alternative, but that some *malin génie* will do you in if he discovers that you *hope* that *S*. Then in one sense you have reason to hope that *S*, but in another sense you have decisive reason not to hope that *S*, and so on.[6] Derek Parfit calls reason for the first sort "object given reasons", since they seem to derive from features of the object of the attitude in question.[7] He calls reasons of the second kind "state-given reasons", since they seem to derive from features that the attitude would have if one were to adopt it. There is however no generally accepted account of how to draw this distinction, and we cannot pursue the general problem here.[8] The crucial point for our purposes is that if we are going to speak of "rationality" at all in this connection, "rational" cannot mean: *rational all things considered.* It must mean, "*rational taking only the object-given reasons into account*." But given the obscurity of this latter notion, we cannot simply rest with this analysis. The proposals to be considered below may be understood as attempts to clarify this

[6] Kavka's Toxin Puzzle presents another instance of this pattern.
[7] "Reasons and Rationality", ref.
[8] Refs to Jacobson and D'Arms, Rabinowicz, Hieronymi.

notion, at least for the special case of the reactive emotions, and so to provide a more informative response to the Cleopatra problem.

3. The Fittingness View.

Since the word "rational" exhibits a crucial ambiguity in this context, let us set it aside and return to our dummy word, "appropriate". There is no doubt that anger would be an appropriate response to Cleo's act. The question is: What is it for a response to be appropriate in this sense?

Here is one possibility. Consider the relation that obtains between laughter and the funny, or fascination and the fascinating, or excitement and the exciting. This relation, sometimes called "fittingness", has a number of suggestive features in common with the relation we have called "appropriateness". It is clearly a normative relation. But it cannot be identified with all things considered rationality or justification. A joke can be riotously funny even though it would be a mistake all things considered to find it funny. Nor can it be identified with any of the more restricted normative relations with which practical philosophy is standardly concerned. An encounter can be exciting even though it would be prudentially irrational or morally wrong to be excited by it. And a moment's reflection suggests that the same is true for appropriateness. An agent can be culpable for an act even though it would be morally wrong or counterproductive to blame him for it. And we can easily imagine cases in which morality or prudence requires blame even though the agent is not in fact blameworthy.[9] This pattern may suggest that we are here in the presence of a single pervasive species of evaluation.

[9] This is presumably how it is with children. We may have good moral and prudential grounds for blame — not just for faux blame, but for the genuine article — even when it is clear that the child is not in fact blameworthy.

Proponents of this approach generally hold that fittingness is irreducible.[10] Indeed, many have proposed to regard it as *the* primitive normative relation, hoping to define the more specific concepts of value theory in terms of it.[11] Whether or not this project succeeds, it does seem plausible at this stage that if there is such a thing as fittingness, it cannot be analyzed in terms of anything more basic.

According to the Fittingness View, an agent is blameworthy just in case resentment (and the rest) would constitute a fitting response to the act in question. The view cannot be excluded at this stage, but from a theoretical point of view it would be somewhat disappointing if we were driven to it.

Recall that our main theoretical ambition in this area is to articulate the principles governing appropriate resentment. Suppose you blame the idiot in the blue Mercedes for cutting you off in traffic and someone says, "Look, you shouldn't blame him. You were in his blind spot. He couldn't see you." This remark is obviously relevant, and if we were to flesh out the example, it might be decisive. Such examples suggest that appropriate resentment is governed by principles. As a first pass in this case we might propose the following: resentment is inappropriate when the act in question was done from blameless ignorance of some fact that bears on its moral permissibility.[12] But of course this may not be quite right, and as theorists we might set ourselves the task of stating the operative principle more exactly. But if appropriateness is fittingness, it would be surprising if principles of this sort were ultimately forthcoming. There are no interesting general principles governing the fittingness of laughter or excitement or interest. It may well be that beauty is properly analyzed in terms of fittingness: a thing is beautiful iff a certain aesthetic response to it would be fitting. But this is plausible in part because (as long critical experience has

[10] Ref. To Gibbard on Ewing, Jacobson and D'Arms.
[11] Wiggins, Jacobson and D'Arms.
[12] See Gideon Rosen, "Culpability and Ignorance", Proceedings of the Aristotelian Society, ref; "Skepticism about Moral Responsibility," Philosophical Perspectives, ref.

shown) there are no interesting principles governing aesthetic response — no general truths of the form: X merits aesthetic appreciation if/only if X is ϕ. So if the Fittingness View is correct, we should not expect to discover general principles governing blameworthiness. We should not expect an articulable doctrine of the excuses. And not only would this be disappointing. As will emerge, we have excellent grounds for thinking that there can be such a doctrine. And to the extent that this optimism is warranted, we have reason to doubt the Fittingness View.

Recall that another ambition is to provide an account of blameworthiness that might furnish rational leverage in disputes over novel putative excuses. If the fittingness view is correct, this is almost certainly a pipe dream. We can know that a joke is funny if amusement would be appropriate, or that a picture is beautiful if it merits aesthetic appreciation. But such knowledge provides no leverage at all for adjudicating particular judgments or putative general principles of comic or aesthetic value. So if we were driven to the Fittingness View, we would probably have to abandon this ambition.

Recall finally that our theoretical aim is not simply to catalog the principles governing the appropriateness of blame. We also aspire to explain and justify those principles. Now suppose that the excuse mentioned above is utterly genuine: If X does A from blameless ignorance of A's wrong-making features, then X is not blameworthy for A. We might then want to know *why* this is true. Why should it be that resentment is in appropriate when an act has this (somewhat arcane) feature? If the Fittingness View is correct, we should not expect answers to this sort of question. In other domains, even when we can identify features that are relevant to the fittingness of a response, there is really no prospect of an account of *why* those features should be relevant. We may suspect that laughter is fitting when a performance exhibits a certain (ineffable)

kind of incongruity. But when we ask *why* laughter should be fitting under such circumstances, we have no idea where to begin. The inquiry is fruitless on its face. If the Fittingness account of culpability is correct, then the prospects for an explanation of the catalog of excuses are dim indeed.

So given our theoretical ambitions, we have reason to *hope* that the Fittingness View is not correct. But of course this is not a serious objection to the view. So for now, at any rate, the Fittingness View remains on the table. The question is whether it is possible to do better.

4. The Fairness View.

It is often said that blame is a form of punishment, or that it is like punishment in certain respects. The crucial point of the analogy is supposed to be that blame and punishment both amount to *moral sanctions*: harsh treatment imposed for the violation of a moral norm. If this is right then it is plausible that that like other sanctions, blame should be governed by distinctively moral norms of fairness and desert. The relevant notions of fairness and desert may be somewhat elusive. But consider plausible principles such as the following:

No one deserves to be sanctioned for conduct if that conduct was not wrong.

It is unfair to sanction someone for violating a norm if through no fault of his own he lacked the capacity to comply with it.

This suggests an alternative proposal according to which an agent is blameworthy for an act iff a certain sort of moral sanction — the psychic sanction of blame — would be fair or deserved. A consideration counts as an excuse on this view iff it entails that psychic sanctions would be unfair in the intended

sense. Call this the Fairness View. The view identified the normative relation we have called "appropriateness" with a form of moral fairness.

Like the Fittingness View, the Fairness View may be a form of primitivism, since it may well turn out that the relevant notions of fairness and desert resist analysis. But it would clearly be a different form of primtivism, since the fittingness norms that govern laughter and excitement are clearly different from the moral norms that the Fairness View invokes. And while we may despair of the prospects for an articulate theory of fittingness, a theory of the conditions under which harsh treatment counts as fair is hardly out of the question. So this approach at least holds out the prospect of principled resolutions of controversial questions about the conditions of blameworthiness, even if those prospects are (at this point) somewhat remote.

The Fairness View has been developed in detail by Jay Wallace[13]. Its main defect, in my view, is that the crucial analogy between blame and punishment is broken-backed. Punishment typically causes pain and misery, and we typically punish in order to cause misery. By contrast, while it is true that resentment may sometimes cause psychic pain, this is hardly an essential or even a typical feature of it. Moreover it seems frankly wrong to suggest that we resent in order to cause pain. (Think about your resentment of the idiot in the blue Mercedes who has just cut you off in traffic.) So even if we grant that overt sanctions are governed by moral norms of fairness in virtue of the fact that they involve the deliberate infliction of psychic pain, or simply because they foreseeably cause such pain, it would not follow that the reactive emotions are similarly governed by such norms. Now in the end I believe they are, and I will sketch a positive argument for this conclusion. But unlike the argument supported by the Fairness View, my argument does not depend on the assumption that blame is a sort of punishment, or that it is governed by norms of fairness because it amounts to a form of sanction.

[13] Responsibility and the Moral Sentiments, op. cit.

5. The Alethic View.

Return to our example. The idiot in the blue Mercedes has just cut you off in traffic. But your were in his blind spot, and this means (we may stipulate) that through no fault of his own, he did not know that by swerving into your lane he would be cutting someone off. Why is this fact relevant to his culpability? Here is a natural first thought. In blaming the idiot you took his action to be an expression of *ill will*. You assumed that either he knew what he was doing, or that he didn't care enough to check his mirrors. You thus construed his act as an expression of a lack of concern for your rights and interests. The excuse is relevant because it shows this assumption to be false.[14] And from the fact that this assumption is false it seems to follow that your resentment was inappropriate, and hence that the idiot is not blameworthy.

This natural thought suggests a theory. Resentment and the other reactive attitudes are emotions, and like many emotions they involve thoughts about their objects. It seems immensely plausible, for example, that when you resent someone for an action, you think of his act as wrong, or as something that he should not have done. And as the example suggests, it is also plausible that resentment involves the thought that the act expresses an objectionable attitude towards those affected: an insufficient degree of concern or respect. These thoughts need not amount to full-blown judgments or beliefs. It is possible to blame yourself for an action (by feeling guilty) even though you know full well that you didn't do anything wrong, or that it was an honest mistake expressing no ill will. In such cases, the moral thoughts implicit in blame are like the cognitively impenetrable appearance of unequal lines in the Müller-Lyer illusion. The

[14] For this account of excuses like ignorance and inadvertence, see Strawson, op. cit.

reactive attitudes, as we shall say, *present* their objects as having certain moral features. The subject of the emotion need not assent to the presentation.[15]

According to the Alethic View, an emotion is appropriate in the relevant sense just in case its ingredient thoughts are true — just in case the emotion presents its object as it really is. We need not suppose for this purpose that emotions are individuated by their ingredient thoughts. We may allow that two emotions, though distinguishable in other ways, may be constituted by the same range of thoughts. The Alethic View would then entail that these emotions have the same "appropriateness" conditions. But it would entail that they were the same emotion.

The Alethic View promises to make straightforward sense of the drama of indictment and excuse. The argument implicit in our example might run as follows:

To resent X for A is to think, among other things, that X expressed ill will in doing A.

The idiot did not express ill will in cutting you off.

Resentment is appropriate iff its ingredient thoughts are true.

So resentment is inappropriate in this case. The idiot is not culpable for his act, wrong thought it was.

The Alethic View holds out the promise of non-trivial explanations of this sort across the board: explanations that help us to understand why the

[15] Nussbaum, *Upheavals of Thought,* identifies emotions with beliefs or judgments and explains such cases as cases of inconsistent belief. I find the appeal to presentations more plausible, but nothing in the present discussion hangs on this.

uncontroversial excuses are excuses and which might even help us to decide whether some new and controversial excuse should be allowed. The question will always come down to this: does the novel putative excuse negate some thought implicit in resentment and the other reactive attitudes? If it does, it is genuine; if not, not.

More generally, the Alethic View promises to furnish a powerful tool for inquiry. On any view, a complete theory of these matters will include an account of the propositional content of the reactive attitudes. It will also include a catalog of excuses. We begin with some insight into both aspects of the theory. Reflection tells us something about what we're thinking when we resent someone for what he's done, and ordinary moral practice serves up confident verdicts about certain putative excuses. But of course we do not begin with anything like a complete account in either area. The Alethic View licenses us to rely on our views about the contents of the reactive attitudes to inform our theory of the excuses, and vice versa. If we think that some thought is implicit in resentment, then the Alethic View entails that any consideration incompatible with that thought is an excuse. And if we think that we have identified an excuse, the View instructs us to find a thought implicit in resentment with which it is incompatible. The fact that the View places such stringent constraints on our evolving theory is not a reason to believe that it is true. It is, however, an excellent reason to take to it seriously.

6. The Naivety Constraint.

The Alethic View has another consequence that will be important in what follows. As we have repeatedly stressed, a complete theory of these matters will specify the conditions of culpability — perhaps positively, by specifying the conditions themselves, and perhaps negatively, by identifying the most fundamental excuses. We may now note that the Alethic View places

significant constraints on any such theory: it must be formulable in naïve terms. More specifically, it must be possible to formulate the theory using concepts that are all available to a naïve subject — a subject capable of resentment but otherwise ignorant of philosophy.[16] The reason is straightforward. Anyone capable of blame must be capable of thinking the thoughts that are constitutive of blame. The Alethic View entails that these thoughts determine the conditions of culpability. This means that if we had an explicit formulation of these thoughts, we could specify the conditions of culpability just by listing them and saying: X is blameworthy for A iff the thoughts on the list hold good for X and A. The resulting account might not be the most perspicuous account of the conditions of culpability. But it would have to be a correct account all the same.

To see that this constraint has teeth, consider an elegant version of the Alethic View due to Nomy Arpaly:

> I take blame not to be an inner version of social sanction or a practice, but a belieflike attitude similar to various kinds of esteem. Blame is not something primarily required or prohibited, like punishment, nor even something that can be appropriate or inappropriate, the way that a brave attitude is appropriate for a soldier. It is first and foremost *warranted* or *unwarranted* in the way that my fear of getting a flu shot is warranted only if flu shots are dangerous to me. … To hold someone blameworthy is not, in itself, to hold that any course of action is appropriate or inappropriate in regard to him, but rather to hold that a certain attitude towards him is epistemically rational: there was ill will; there was a wrong act, thus blame is warranted. In this way, on my view, blame is analogous to holding someone to be a bad businessman or a lousy artist.[17]

[16] The naïve subject need not have words for these concepts, if he can think without words.
[17] *Unprincipled Virtue*, (ref.) 172-3.

This passage suggests a simple view: to blame X for A is simply to judge (or perhaps to think) that A was wrong and that X displayed ill will in doing it. For the act to be blameworthy just is for these thoughts to be true.[18]

This simple theory plausibly satisfies the naivety constraint. It is always hard to know how to impute thoughts to people who may not have the words to express them. But it is plausible that anyone capable of genuine blame must be capable of regarding an act as an *offense* — as a wrong that expresses an objectionable attitude towards those affected. Someone who could not think of an act as an offense could not see the point of exculpating remark like "He didn't mean to" or "He had every right to do it". And if someone cannot see the point of such remarks, it is unclear with his negative response deserves to be called "blame", as distinct from some more primitive form of anger or contempt.

The trouble, of course, is that the simple theory is mistaken. A child can act wrongly and with ill will. A demented adult can be cruel, not just inadvertently but out of malice. And yet in many such cases, blame is obviously inappropriate.[19]

The Alethic View thus forces us to find an ingredient thought in the reactive attitudes that fails to obtain in these case — a thought negated by (what the lawyers might call) infancy and mental disease or defect. One might be tempted to meet the challenge directly by supposing that blame is partially constituted by the thought that *the agent was a sane adult when he acted*. But on reflection, this won't do. We can easily imagine someone who blames as we do but who lacks the concept of mental illness. (Suppose that where he comes

[18] Arpaly should not say that blame is warranted when the underlying thoughts are epistemically rational, since they may be rational but false, in which case the agent is not blameworthy. The examples suggest that this is a slip, and that she intends the Alethic View.

[19] Arpaly is of course aware of this difficulty. For her discussion, see *Unprincipled Virtue*, ch. 6.

from there has never been any such thing, and that it has never occurred to him imagine the possibility.) More fancifully, we can imagine someone who blames as we do but who lacks the concepts of adulthood and childhood. Such people would lack the capacity to think of someone as a sane adult. But this incapacity would not prevent them from blaming one another for their transgressions. Any plausible version of the Alethic View will therefore have it that when we blame we think a thought that is in fact incompatible with certain forms of infancy and insanity, and perhaps also with extreme emotional disturbance, involuntary intoxication, hypnotism, somnambulism and the rest, even though that thought does not involve explicit reference to these categories. Hence a pressing question for the Alethic View: What might this further thought be?

7. The Capacity for Rational Self-Control.

Wallace has argued persuasively that what unifies this class of excuses is that in each case the agent lacks, to some substantial degree, what he calls "the capacity for reflective self-control". This is really a complex of capacities that includes the capacity to step back from one's impulses in order to ask whether one has reason to act on them, the capacity to address this question by deliberating correctly about what to do, and the capacity to control one's conduct in light of the upshot of such deliberation. Wallace's plausible claim is that small children, the insane, those who act under duress, and so on, are to be excused to the extent that they lack this complex capacity.

Can the Alethic View incorporate this insight? That depends on whether we can plausibly attribute thoughts about the capacity for reflective self-control to the naïve subject. Is it plausible that just as the naïve subject of blame must be in a position to think of an act as an offense, he must also be in a position to think of its author as possessing this complex of rational powers? This may seem implausible. Consider the naïve subject who is fuming at the idiot who has just

cut him off in traffic. His attention is focused on the act, and perhaps on the attitude he takes it to express. Is it plausible that he must also be thinking of the driver as someone who possesses the capacity for rational self-control? To bring out the implausibility of this suggestion, consider the naïve subject in a world in which every being capable of thought and action — every rational agent — has always possessed this capacity. In such a world, it may never have occurred to anyone to distinguish the capacity for rational self-control from the capacity for action more generally. The proponent of the Alethic View who wants to borrow Wallace's insight must apparently hold that even in *this* world, blame presents its object as capable of rational self-control. This strikes me as implausible. But since it remains somewhat unclear what constrains our attribution of thoughts to the naïve subject, perhaps we should not place too much weight on this intuition.

If we waive this objection, the Alethic View may take the following form: To resent X for A involves (so far as thought is concerned) the thought that A was wrong, the thought that X expressed ill will in doing it, and the thought that at the time X possessed the capacity for reflective self-control. Since blame is appropriate iff its constituent thoughts are true, it would follow that X is culpable for A iff these three conditions are satisfied. We would then have an explanation for why it is that infancy, insanity and the other forms of incapacity excuse when they do, an explanation whose main premises are a hypothesis about the content of the reactive attitudes and a statement of the Alethic View itself. We would also have a strategy for assessing putative novel excuses: the question in every case will be: Does the putative excuse negate one of the three thoughts that constitute the reactive emotions?

8. A Complication.

This is an elegant package, but it cannot be quite right as it stands. Suppose that Fred maliciously shoves an old lady into the gutter, and suppose it emerges

on closer inspection that he was out of his mind when he did it. Perhaps he knew at some level that his act was wrong. Nonetheless, he was consumed with a mad rage, and as we may stipulate, quite incapable of steering by this still small voice. The theory now on the table entails that Fred is off the hook. But in fact this is too quick. Suppose that Fred was out of his mind at the time only because he had deliberately ingested a drug designed to boost the aggressive impulses while destroying the capacity for self-control. (Think of him as a method actor eager to know how it really feels to lose control.) Well, then it seems that Fred is back on the hook.

It is easy to state the principle that underlies this verdict. Incapacity is an excuse, *but only when it is non-culpable*.[20] Fred is culpable for his incapacity, since it was the foreseeable upshot of a culpable reckless act.[21] And in that sort of case, incapacity is no excuse.

If the Alethic View is to absorb the point it must complicate its account of resentment. The most straightforward approach would be to build the above diagnosis into the content of the reactive attitudes. We would then have to imagine that when the naïve subject resents someone for an action, he thinks of the agent as someone who either possesses the capacity for reflective self-control or else culpably fails to possess it. The Alethic theorist who pursues this route is thus led to say that blame is constituted in part by thoughts about blameworthiness, which is to say that it is constituted in part by thoughts about blame itself and about the norms that govern it.

This is not by itself a decisive objection to the view. It would be implausible to suggest that the emotions are in general self-referential. But some of the higher emotions may well be. The trouble is rather that at this point the naivety

[20] This parallels the observation, noted earlier, that ignorance of wrong-making features is an excuse, but only when it is blameless.
[21] For doubts about whether real cases of "reckless" incapacity satisfy this condition, see my "Skepticism about Moral Responsibility" ref.

constraint has almost certainly been flouted. It is not implausible that the naïve subject should be in a position to think about blame, or even about blameworthiness. If he is at all familiar with the business of offering excuses, he must have some sense that blame can be appropriate or inappropriate, and it is arguable that his sentiment would not amount to blame if he were utterly unfamiliar with the practice of excusing. The objection is rather that it is simply crazy to suppose that when the naïve subject is fuming at the idiot who has just cut him off in traffic, his thought must somehow encompass the possibility that its target culpably lacks the capacity for rational-self-control. If the naivety constraint has any teeth at all, it must rule out this proposal.[22]

The proposal is phenomenologically improbable in any case. Just think about what you are thinking when you are in the throes of indignation. Does anything in your thought correspond to the disjunctive condition we have been considering? Is it plausible that your thought explicitly looks forward, as it were, to the possibility of voluntary intoxication (or culpable incapacity) and excludes it as a potential excuse? This seems unlikely.

10. An Alternative Proposal and a Logical Problem

At this point I have sometimes heard the following suggestion. The naïve subject need not entertain the disjunctive thought presently under discussion. He need not think about the capacity for reflective self-control at all. If he possesses the concept of an excuse, as we have supposed he does, that is enough. The naïve subject in the throes of blame need only think: That was

[22] Strictly speaking, we should distinguish the basic naivety constraint — which requires that the naïve subject who is capable of blame must possess the concepts that figure in the thoughts constitutive of blame — from a related constraint. This related constraint requires that the thoughts constitutive of blame are operative even when the naïve subject goes in for the most elemental and unreflective forms of resentment. The objection in the text is that the revised version of the Alethic View flouts this latter constraint.

wrong; it manifests ill will; and *he has no excuse.* Equivalently, he need only think: That was an offense *and he's blameworthy for it.*

This proposal is much more plausible as a matter of phenomenology. It is also compatible with the naivety constraint. It is, however, liable to a decisive objection. The fundamental premise of the Alethic View is that when an act is blameworthy, it is blameworthy *because* the thoughts that are constitutive of blame are true of it. On the present proposal, this amounts to the claim that X is blameworthy for A iff A was wrong, A manifests ill will, and *these three thoughts are all true*. But this is hopeless. The proposal entails that the content of an episode of blame is an *ungrounded* proposition. It is a variant so-called *truth-teller* ("This very sentence is true.") of the form:

Sentence S: *P* & *Q* & *'S'* is true.

But as is well known, such statements lack determinate truth conditions. When *P* and *Q* are both true, we may consistently suppose that *S* is true; but we may also consistently suppose that it is false, or that it lacks a truth-value. The present proposal thus entails that the thoughts implicit in blame will never be determinately true, and hence that no one will ever be determinately blameworthy for his actions. Clearly the account has run off the rails.

11. A final proposal.

I will mention one more strategy for the proponent of the Alethic View. Recall the idea mentioned earlier that blame is a sanction akin to punishment and therefore governed by moral norms of fairness. As noted, the analogy is in certain ways far-fetched. If I am indignant at the hypocritical pandering of a congressman, there is no real sense in which my sentiment constitutes an effort (much less, a successful effort) to make him suffer. When I am fuming at the idiot who has cut me off in traffic, I may be animated by violent fantasies but my

thought is not itself an act of violence. Of course, insofar as each of us as an interesting in being liked or well regarded, my resentment may constitute an objective set back to his interests. But as Pamela Hieronymi has stressed, it does not follow from the fact that an attitude of mine sets back your interests in this way that it is governed by non-trivial norms of norms of fairness.[23] If I find you ugly or untrustworthy and you have an interest in being well regarded, then my attitude sets back your interest. But if you *are* ugly or untrustworthy, then there is no further question about whether it is fair for me to adopt the attitude. In such cases, it is a stretch to think of my attitude as something that I do to you — as a form of *treatment*. It is certainly not something I do in order to harm you or with the expectation that it will harm you. So negative sentiments are not in general a form of sanction, and they are not in general governed by non-trivial norms of fairness. The suggestion, implicit in Hieronymi and others, is that we should think of the reactive attitudes in a similar spirit and so reject the analogy with punishment that underlies the Fairness View.

Nonetheless, the analogy is not altogether empty. There is a difference between finding someone ugly or untrustworthy on the one hand, and resenting him, on the other. Negative assessments of the first sort may be accompanied by pity and solicitude. The reactive emotions, by contrast, are essentially *hostile* thoughts. In what does this distinctive hostility consist? One possibility is that the hostile sentiments involve a desire to harm. This is probably the right account of the hostility implicit in the most primitive forms of anger: the sort of anger a dog can feel. But in the case of the reactive attitudes it seems more plausible to say that their inherent hostility consists at least in part in the thought that the target *should* suffer, or that the target *deserves* to suffer, for what he's done. Let me suggest a somewhat more nuanced version of this idea.

The reactive emotions are not sanctions, but they are naturally *expressed* by sanctions. When you blame someone for what he's done, one natural

[23] Hieronymi, "Controlling Attitudes", also Arpaly, op. cit.

expression of your emotion is to yell at him or to hit him, all with an eye, in part, to making him suffer for what he's done. (Of course we civilized adults normally inhibit the tendency to express our reactive sentiments in these ways.) It is a plausible speculation that we develop the capacity to think about the reactive emotions, not just by noticing them in ourselves but by noticing them in others. And when we notice them in others we do so by noticing their manifestation in overt moral sanctions of this sort.

The proponent of the Alethic View might then suggest that among the thoughts that constitute blame are thoughts about the fairness of moral sanctions. We might say that to blame a person for an act is to think of that act as an offense for which overt sanctions would be fair or deserved. Blameless incapacity would then constitute an excuse, not because blame itself attributes the capacity for reflective self-control or its culpable absence to the agent — as we've seen, that thought is to recherché to figure in ordinary resentment — but rather because it involves the thought that the agent deserves to be sanctioned for what he's done. The case for recognizing blameless incapacity as an excuse would then have to proceed, as in the Fairness View, by appeal to a substantive moral principle, viz., that someone blameless lacks the capacity for reflective self-control it would be unfair to sanction him for conduct that is somehow the result of that incapacity.

Is the proposal compatible with the naivety constraint? As always, it is hard to say. But it seems plausible that it is. As noted, it is likely that anyone capable of blame will be familiar with its overt manifestations, and also (we may imagine) with the fact that these overt manifestations are governed by norms. He may not have a refined vocabulary for articulating these norms. But it is not crazy to suppose that anyone capable of blame (as distinct from simple anger) must have some grip on the distinction between excuses — considerations that tend to show that sanctions would be undeserved — and the various other considerations that might be invoked against sanctioning behavior. If he does

then he has the raw materials for thinking the thoughts that are constitutive of blame on the present view.

This is the best I can do for the Alethic View. If the proposal is on the right track, then a full development of the view will overlap considerably with a parallel development of the Fairness View. If the Alethic View is to be preferred, that is because it explains why considerations of fairness and desert should be relevant to the appropriateness of blame without appealing to the dubious assumption that blame is itself a form of sanction. Both views are of course obliged to say more about the relevant concepts of fairness and desert. But there is no reason at this stage to doubt that this story can be told.

We may note one striking consequence of this version of the Alethic View. It is often supposed that questions about the fairness of overt sanctions are derivative, in the sense that the answers depend, in part, on a prior judgment about the moral culpability of the agent. It is common, for example, for a theorist to say that it would be wrong to sanction the insane *because the insane are not morally culpable for their actions*. On the present view, this is backwards. It would be more accurate to say that the insane are not morally culpable for their actions *because it would unfair to punish them* or to sanction them in other ways.[24]

12. A Puzzle about Forgiveness. (A digression, even rougher than the rest.)

The Alethic View depends on the idea that the reactive emotions are partially constituted by certain thoughts. So far we have been neutral as to whether these emotions are *exhausted* these thoughts. The Alethic View is most clearly at home in a cognitivist or judgmentalist framework according to which the emotions

[24] This is not to say that considerations of moral culpability are not part of the justificatory basis of state punishment. State punishment is a very special case. The sanctions at issue in the account of resentment are informal personal and social sanctions.

are to be identified with constellations of belief-like representations. After all, if the reactive emotions have non-judgmental components — desires, wishes, motivational tendencies, perhaps mere affect — one might wonder why the appropriateness conditions for these emotions should be determined entirely by their judgmental constituents. Certain non-judgmental states — desires in particular — are governed by relevant norms. So why should the appropriateness of a composite state which contains a desire as a constituent depend entirely on the appropriateness conditions of only *some* of its components? This is hardly a decisive argument. The point is to suggest that the Alethic View would follow naturally from the cognitivist assumption that emotions are exhausted by their component thoughts, whereas the View sits somewhat uneasily with alternative views of the emotions.

There are of course many objections to the cognitivist view as a general theory of the emotions. My sense is that most of them have been answered. There is, however, a problem that arises in connection with the reactive emotions in particular.

Suppose the idiot is in fact blameworthy for his bad driving. At first you resent him for it — appropriately — but then, after a decent interval, you forgive him. What exactly happens in this transition? It is a commonplace that forgiveness is distinct both from forgetting about the wrong and from excusing or condoning it. Forgiveness is forgiveness only so long as one continues to think of the agent as genuinely culpable for the act. In Butler's phrase, to forgive is to *forswear* resentment; it is not to abandon the judgment that the act is resentment-worthy. And this raises a question for the thesis that resentment consists in a constellation of belief-like thoughts. If resentment consists in the thought that the act was wrong and that sanctions would constitute an appropriate response to it, in what sense can one forswear resentment while retaining these judgments?

I can think of no plausible answer to this question. It seems to me quite likely that insofar as resentment is a hostile sentiment, it involves an ingredient over and above the judgments we have been discussing. This further ingredient might be a desire to harm; it might be a dim but gratifying fantasy of harm. If this is right then we may say that in forgiveness one forswears resentment by forswearing hostility in this sense. And this is evidently something that we can do while retaining the judgment that the act was an offense for which sanctions would have been appropriate.

But then the question will arise: Why doesn't this non-judgmental component of resentment bear on the appropriateness conditions of the emotion? And here a natural answer suggests itself. It may be that the desires or fantasies that constitute the hostile aspect of resentment are subject to assessment as appropriate or inappropriate. But if so, it is plausible that these non-judgmental components of the response are appropriate *precisely when* the corresponding judgment of desert is true. It may make sense to respond to the idiot's bad driving with a desire to see him suffer. But it is plausible that this desire to sanction X for A is appropriate precisely when such sanctions would not be undeserved. If this is right, then we can explain why the appropriateness conditions for the reactive emotions are given entirely by the truth-conditions of their ingredient thoughts. It is not that the emotions are exhausted by these thoughts. It is rather that the non-judgmental component of the emotion — the hostile element — makes a *redundant* contribution to the appropriateness conditions of the emotion.

12. Conclusion.

This paper has been concerned with three questions: (a) What is it for X to be morally blameworthy for A? (b) What are the conditions under which X is blameworthy for A? And (c) *Why* are those the conditions under which X is blameworthy for A? The Alethic View is an intriguing answer to the first question.

We can think of it as supplying answer to the question that we pressed on Gibbard: What sense must we supply for the words “rational” or “warranted” if we are to defend the claim that an agent is blameworthy for an act just in case resentment and indignation are rational or warranted? The Alethic View says: A response is rational or warranted in the intended sense when its ingredient thoughts are true. The account has many virtues. The most important is that it holds out the possibility of substantive answers to questions (b) and (c). The view has the further virtue of vindicating the fundamental claim of the Fairness View — that considerations of moral desert have some bearing on questions of blameworthiness — without assuming, what is implausible, that moral blame is itself a form of sanction governed by norms of fairness. It remains to be seen, however, whether the view is tenable in the end. A complete defense would require a perspicuous formulation of the Naivety Constraint and a demonstration that the best versions of the Alethic View are consistent with it. We also require a more careful analysis of the hostile element in resentment if we are to be confident that this element makes, at best, a redundant contribution to the appropriateness conditions of the reactive attitudes. At this point I hope only to have made it plausible that the Alethic View merits further consideration.