CHAPTER 1

The Motivational Theory of Guilt (and Its Implications for Responsibility)

Justin D'Arms and Daniel Jacobson

The Strawsonian approach to responsibility tries to explain what it is to be morally responsible for one's actions in terms of being an appropriate object of the reactive attitudes (see Strawson, 1962).[1] In order to succeed, the approach must first explain what the relevant attitudes are and what is meant by appropriateness. Although there are both negative and positive reactive attitudes, corresponding to blame and praise, most of the discussion following Strawson focuses on the negative side. It can therefore only hope to capture blameworthiness, not responsibility in general, since to be morally responsible in a good (or neutral) way is surely not to be the appropriate object of a negative attitude. We, too, will focus on blameworthiness, which Strawsonians hope will provide the foundation for a general theory. This chapter develops and answers an important challenge to any such account of responsibility, whatever the reactive attitudes to which it appeals. Our discussion centers on guilt, for reasons to be explained, and hence specifically concerns self-blame. A similar problem arises for other-directed blame, which will require an analogous solution.

The challenge facing the Strawsonian project also faces the sentimentalist project we have been developing for some time, and we will suggest that the same solution applies to both cases. *Sentimentalism*, as we understand it, refers to those views that explain (at least some) values in terms of the emotions; and our own view, *rational sentimentalism*, does so specifically in terms of the fittingness of emotions – or, equivalently, of what merits them – where merit and fit are understood to be notions of correctness. We have argued that considerable confusion arises from the failure to differentiate between fittingness and other forms of appropriateness.[2]

[1] There are other ways to read Strawson's classic paper, but this is what we shall mean in referring to the Strawsonian tradition. McKenna (2012), Rosen (2015), Shoemaker (2017), and Wallace (1994), among others, are all Strawsonians in this sense.

[2] See D'Arms and Jacobson (2000) for more on differentiating such notions of appropriateness.

https://doi.org/10.1017/9781009179263.002 Published online by Cambridge University Press

An influential challenge presented by Philippa Foot observes that sentimentalist explanations are informative only if the emotions they appeal to do not already include the evaluative concept they attempt to explain. In her view, sentimentalism fails that challenge because "the explanation of the thought comes into the description of the feeling, not the other way round" (1978, p. 76). Foot adopts a cognitivist theory of the emotions, in which they are type-identified by some constitutive thought necessary for having the emotion. In order to be proud of something, for example, one must believe it to be splendid and one's own. As she puts it: "I do not mean, of course, that one would be illogical in feeling pride towards something one did not believe to be in some way splendid and in some way one's own, but that the concept of pride does not allow us to talk like that" (1978, p. 76).

According to Foot's challenge, sentimentalism gets the order of explanation wrong. Emotions are to be explained in terms of values, not values in terms of the emotions. If to be prideworthy is to merit pride, and pride is even partly constituted by the thought that something is splendid and mine, then it seems to follow that for something to be prideworthy is just for it to be splendid and mine.[3] But if the prideworthy can be understood via a pride-independent notion of *splendid and mine*, then sentimentalism would be otiose: pride drops out of the explanation of the prideworthy. A distinct but related problem is suggested by Foot's cognitivist claim that the order of explanation goes from the evaluative concept to the emotion, "not the other way round." Some sentimentalists propose to adopt both directions of explanation in an overtly circular fashion: the value gets explained in terms of the response, which in turn gets explained in terms of the value (see Wiggins, 1987). We are skeptical of the claim that such an explanation is not viciously circular; at any rate, we do not think that it can explain much.

The same issues arise for the Strawsonian account of responsibility. If blameworthiness should be understood via some reactive attitude whose content can be given in terms of concepts that are attitude independent, then the attitude seems to drop out of the explanation. If to be blameworthy is to have violated a requirement of respect, for instance, then – even if there is some reactive attitude that involves the thought that someone has disrespected you – the attitude seems inessential to this account of

[3] At any rate, this is so if fittingness is tantamount to the truth of the emotion's constitutive thought. Indeed, cognitivism's ability to explain fittingness in this straightforward way is one of its features.

blameworthiness (cf. Graham, 2014). Such an explanation renders the attitude otiose.[4] Yet, if blameworthiness must be explained in terms of a reactive attitude that is even partly constituted by a thought containing concepts such as *blameworthiness* or *responsibility*, then the explanation would be rendered circular.

We suggest that Foot's challenge sets the ground rules for a successful Strawsonian account. The reactive attitude to which it appeals must meet two conditions: (1) *Priority*. The attitude's content must not be capable of being given in wholly response-independent terms, or the attitude will drop out as otiose. And it must not be given in terms of responsibility or any concept that presupposes responsibility, on pain of circularity. (2) *Rational assessability*. The attitude must be amenable to assessment of its appropriateness in the relevant respect, such that it is appropriate to respond that way specifically to blameworthy (or otherwise responsible) action. The trouble is that the most straightforward way to meet the second condition seems to presuppose a cognitivist theory of emotion that cannot meet the first condition.

Philosophers tend to conceive of reactive attitudes as propositional attitudes, and to characterize them in terms of certain thoughts, beliefs, or judgments necessary for their possession. This approach risks violating the priority condition. In our view, a core class of what we term natural emotions are the most promising candidates for a Strawsonian account, because they have a psychological character that is independent of the concepts the account tries to explain. Many of Strawson's examples of reactive attitudes are emotions, including indignation, resentment, and guilt – but these emotions differ in one crucial respect. If indignation and resentment are second- and third-personal attitudes whose content involves the notion of wrongness, as is often claimed, then Foot's challenge looms, and it threatens the priority of these attitudes. If wrongness can be understood in wholly response-independent terms – say as what violates the categorical imperative – then the appropriateness of indignation and resentment drops out; it does not contribute to the account of blameworthiness. But if wrongness must be understood even partly

[4] One might be tempted to resist this conclusion by appeal to a distinction between the concept and the property of blameworthiness. It might be said that our concept of blameworthiness is response dependent, involving an essential appeal to some reactive attitude, even if the property of blameworthiness is a response-independent one such as *having engaged in disrespectful behavior*. But this would not vindicate a Strawsonian approach. Whatever one says about the metaphysics of properties and about the conditions of blameworthiness, it is crucial to a genuinely response-dependent approach to responsibility that the reactive attitudes figure in the explanation of why the conditions are as they are, and why this particular property has the significance it has.

in terms of moral responsibility (or blameworthiness), then that would render the account circular.[5]

We have argued elsewhere that resentment and indignation are best understood as *cognitive sharpenings* of anger – a subclass of anger instances that are defined in part by including some thought involving a moral complaint (D'Arms & Jacobson, 2003). Roughly, they involve being angry with someone over her wrongdoing. This suggests that neither of these other-directed reactive attitudes is well positioned to meet the priority condition. David Shoemaker's (2017) recent development of a Strawsonian theory proposes instead that to be blameworthy is to be a fitting target of anger in general. Anger has the advantage of being a paradigm of the core class of emotions that plausibly satisfy the priority condition. But there are difficulties with this suggestion as well. Anger is coarse-grained in some respects, and it is controversial whether its conditions of fittingness match those for (negative) moral responsibility.

The first concern is that there may be a variety of anger, which Shoemaker calls *goal-frustration anger*, that can be fitting without anyone being blameworthy.[6] Another concern is that some actions seem to be suitable targets of self-blaming responses but do not merit the anger of others – cases where one does what one should do, all things considered, but in doing so betrays someone to whom one has special obligations. If so, then the form of blameworthiness that captures responsibility might be better modeled on self-blame than on the blame of others. Finally, Andreas Carlsson (2017) has argued that guilt is uniquely positioned to explain and justify why agents are blameworthy only for what they directly or indirectly control. All of these issues are complex, and they deserve attention in their own right that we cannot offer here. For present purposes, we simply note them as reasons to think that although anger and its cognates have been discussed more and received more of Strawson's attention, there are substantial advantages to a Strawsonian approach that is focused on guilt and, hence, on self-blame.

[5] This poses an interesting problem for Gideon Rosen's *alethic* view. If the thoughts he claims to be integral to resentment can be understood in response-independent terms, and blameworthiness is simply the truth of those thoughts, then resentment would be otiose. Blameworthiness would cease to be response dependent in anything like the way Strawson suggests. It is unclear to us whether Rosen (2015) accepts the first part of the antecedent – the evidence seems equivocal. If he does, he may yet think that he has an answer to the challenge of otioseness, insofar as resentment explains why the conditions of responsibility are as they are. That is a point he makes explicitly. But we think that this explanation would be substantially undermined if the content of resentment can be fully captured in terms of thoughts that are response independent. We cannot pursue that issue further here.

[6] Shoemaker (2018) worries about this possibility and tries to distinguish this sort of anger from what he terms *blaming anger* without circularity.

Guilt, too, must answer Foot's challenge, since its content is also often held to be constituted by thoughts about wrongdoing. We have a theory of guilt to offer, however, which proves helpful because it is suited to play the right sort of role in this dictum:

> (*) For A to be blameworthy for *x* is for it to be appropriate for A to feel guilt for *x*.

We do not here aspire to defend an account of responsibility or blameworthiness on the basis of appropriate guilt, but to develop the building blocks of such an account. Our main contribution is to offer a theory of guilt that can satisfy the priority condition because it is grounded in a sentimentalist-friendly theory of the natural emotions. We can only sketch this motivational theory of emotion here, though we develop it in detail elsewhere. We will then offer some reasons for thinking that appropriateness should be understood as a matter of fittingness rather than some other normative notion. The accounts of guilt and appropriateness we put forward flesh out (*) in a way that avoids problems besetting other Strawsonian accounts. They provide the most promising way to develop a theory of blameworthiness grounded in self-blame.

1 The Motivational Theory of Natural Emotions

We reject the cognitivist theory of emotion, understood as those views that make some constitutive thought a necessary condition for having the emotion and use that thought to type-identify the emotions.[7] As Martha Nussbaum claims: "It seems necessary to put the thought into the definition of the emotion itself. Otherwise, we seem to have no good way of making the requisite discriminations among emotion types" (2001, p. 30). In our view, the putatively response-independent thoughts that cognitivists use to type-identify emotions are either subject to manifold counterexamples or else must become tacitly response dependent. "This is splendid and mine," for example, can be held of many things that are not prideworthy; to take just one example, consider your winning a lottery ticket. Although it is both splendid and yours, it does not seem to merit your pride. Since Foot identifies this thought as a necessary condition for being proud of

[7] This is not to claim that emotions are mere feelings with no cognitive aspect, or to deny that there are conceptions of what it is to have a thought (e.g., "this is dangerous") or to possess a concept (*danger*) such that they can be attributed to an agent simply by virtue of his having an emotion (fear). Such interpretivist views are compatible with the priority thesis, however, unlike traditional forms of cognitivism that challenge sentimentalist and Strawsonian theories.

something, not a sufficient one, it is open to her to elaborate further on the thought – though she never suggests that she sees any need to do so. But in order for her view to belie the sentimentalist order of explanation, as she claims, the additional content must not be pride-dependent. She cannot explain away the lottery ticket example by saying that in order to be pride-worthy, something must be *splendid and mine in the pride-y way*.

Other reasons to reject the cognitivist theory have to do with the nature of emotional motivation and with problems concerning how to adjudicate disputes between cognitivists over the content of these constitutive thoughts. These questions ought to be primarily empirical, but cognitivism seems to make them matters of semantics or conceptual analysis. Moreover, cognitivism has an inadequate explanation of important phenomena such as *emotional recalcitrance* (where an agent has an emotion that is unfitting by his own lights) and *acting without thinking* (where an agent in the throes of an emotion acts on its goal in ways contrary to her ends and sometimes pursues predictably bad means for achieving even the goal of the emotion itself). A motivational theory can do better. Its compatibility with sentimentalism and the Strawsonian approach to responsibility is not the reason to accept the theory so much as a felicitous implication of its acceptance.

Our motivational theory does not attempt to capture all the states commonly called emotions, let alone every affect-laden attitude, but focuses on what we term the *natural emotions*. The natural emotions are pan-cultural psychological kinds that figure in the explanation of various familiar phenomena that would otherwise be mysterious. We are not claiming that everything commonly called an emotion counts as a psychological kind or that only these states should be called emotions. Rather, we use this term to differentiate this core class from cognitive sharpenings (like resentment, as opposed to anger) and from a broad class of affect-laden attitudes (such as love and grief) that the theory does not purport to capture. The natural emotions – which we hereafter will refer to simply as emotions – are goal-directed states characterized by specific *action tendencies*: urgent motivations toward certain actions that are especially direct ways to satisfy the emotion's generic goal, in paradigmatic circumstances.

The goal of fear is threat avoidance, for example, but the state of fear prejudices the means taken to avoid a threat. It favors the most direct and urgent goal-directed actions, such as fleeing. Those threats that are best avoided by calm negotiation or through complex mental calculation still cause fear. Though it may be possible to take these better means despite being afraid, fear impedes its own goal in such cases, because its

action tendency must be overcome in order for the threat to be avoided. Moreover, people in a state of fear are often inhibited in their ability to pursue ends more important to them than avoiding the feared threat – and similarly for other natural emotions. These are respects in which the emotions are discontinuous with practical reasoning, and this is the kernel of truth in the clichéd (and exaggerated) opposition between emotion and reason. Although the motivational aspect of emotions is central to their function, emotions are syndromes that are also typically characterized by other things, including feeling, selective attention, typical elicitors and palliators, bodily changes, and thoughts.

The motivational theory can explain emotional recalcitrance, because it takes the emotions to be discrete motivational systems that are partially encapsulated and, hence, not reliably responsive to certain beliefs and ends that may be contrary to them. The self-aware phobic who is afraid of flying judges it to be less dangerous than many activities she engages in without fear; yet she is disposed to be afraid of flying, nonetheless. Notice how implausible it is to think that she makes conflicting judgments about the safety of flying, given her calm attitude toward the prospect of other people – even those she loves – flying. If she has contradictory beliefs, those are specific to her own flying. The introduction of conflicting beliefs in some such cases seems like a desperate attempt to salvage a theory.[8] Much better to say that her fear motivates her to direct and urgent means of avoiding what it appraises, contrary to her judgment, as dangerous. The motivational theory can offer a similar explanation of acting without thinking. Agents in the grip of an emotional bout are motivated to pursue the generic goal of their emotion in the most direct and urgent ways, regardless of whether these are the best ways to pursue the goal and whether this is the most important goal to pursue.

While various aspects of the motivational theory require further explication, some of them are not crucial for present purposes. What is important here is that it offers a way of understanding the emotions on which they are well suited to satisfy the priority condition. The question is whether there is such a natural emotion that is a likely candidate to play the lead role in a Strawsonian account of blameworthiness. We will argue that guilt is such a state. It is a psychological kind, open to

[8] Although it is possible for cognitivists to hold that agents in the grip of a recalcitrant emotion have contradictory beliefs or conflicting thoughts, those forms of the theory strong enough to undermine the priority thesis have no explanation for why such conflicts persist after their recognition, as ordinary cases of conflicting belief do not – that is, for why they are so recalcitrant to considered judgment.

empirical investigation. It is not even partly constituted by a thought of blameworthiness or by some emotion-independent thought that can explain blameworthiness without appealing to a reactive attitude.

Any theory that attempts to capture blameworthiness in terms of guilt will need a normative component, since it is obviously implausible to understand the blameworthy as whatever actually makes people feel guilty. The fact that someone feels "survivor guilt" over being the only one to survive some catastrophe – assuming for the sake of argument that survivor guilt is a genuine phenomenon and is genuinely a form of guilt, as seems plausible – must not entail that she is blameworthy for surviving. Similarly, for something to be blameworthy is not for it to elicit guilt but for it to make guilt in some sense appropriate. Rational sentimentalism takes the relevant sense of appropriateness to be fittingness. We will assume this position for now and defend it (briefly) later. The question is how to capture what it is for an emotion to be fitting, consistent with the priority criterion. How can one give standards of fittingness for the emotions without appealing to the truth of some constitutive thought? We defend a proposal for how to get fittingness without cognitivism elsewhere, which we can only sketch here (see D'Arms & Jacobson, forthcoming).

Begin with an *empirical* characterization of the general emotional syndrome: the cluster of feelings, patterns of attention, typical elicitors and palliators, characteristic thoughts, and especially the motivational role occurring in paradigmatic episodes of the emotion kind. In light of this data, give an *interpretation* into language of how someone in the grip of such an emotion appraises its object as specifically good or bad. Appraisals in this sense are not constitutive thoughts or components of emotion, but ways of understanding how the emotion as a whole evaluates its object. Any gloss into language will be imperfect and can at most help to point in the direction of the distinctive way that the emotion appraises its object. Since these emotional appraisals are derived from the emotion holistically, including its motivational element, they must be understood as response dependent – even if their terms have response-independent senses in ordinary language. In deciding whether the gloss applies in any given case, one must understand it in a way that is informed by the emotion whose appraisal it attempts to articulate. A minimal condition of adequacy on such a gloss is that it rings true to those who have experienced the emotion. Consider the case of fear and danger.

An empirical characterization of fear favors the suggestion that it should be interpreted as appraising its object as dangerous, for example; this makes

sense of how fear engages with its object – as something to be avoided directly and urgently. Notice too that the manner in which a feared object is to be avoided differs from the way that disgust motivates avoidance. It can be enjoyable to observe something fearsome from safety, whereas one typically wants to avoid perceiving the disgusting. The claim that fear concerns danger is not a surprising suggestion, of course, though interpretive matters are subtler in other cases. What is distinctive about our approach is *how* it understands the claim that fear is about danger: not as a response-independent thought one must have in order to count as afraid, but rather as an effort to articulate the distinctive emotional appraisal involved in the combination of feelings, goals, and action tendencies of fear.

Yet one might be puzzled about how our claim that fear appraises its object as dangerous differs from the cognitivist claim that fear includes a thought about danger. The difference depends on what is meant by saying that fear is (at least partly) constituted by such a belief or thought. We reject a specific and substantive thesis, articulated by Foot and embraced by other cognitivists, which threatens sentimentalist and Strawsonian accounts by violating the priority condition. This is the thesis that emotion types are individuated by some constitutive thought or defining proposition – something explicable independent of other aspects of the emotion, in particular its motivational component, which provides a necessary condition on being in the state. If that were true, then sentimentalism would be otiose; however, these supposedly constitutive propositions are either subject to manifold counterexamples, such as *splendid and mine* with pride, or have to be understood as tacitly response-dependent (see Deigh, 1994; Scarantino, 2010).[9]

On the other hand, if the claim that fear involves a thought of danger does not import these traditional cognitivist commitments, it might be compatible with our view. In particular, if thoughts of danger are attributed to the agent simply on the basis of the fact that she is afraid, then the concept of *dangerous* being imputed can be granted to be tacitly response-dependent. In which case, such a thought or construal may not differ substantively from our notion of emotional appraisal. We find our terminology more perspicuous for making the crucial point, which is that this proposal is compatible with the priority condition and, hence, does not threaten a sentimentalist or Strawsonian account.

[9] Although this is our central and novel objection to the cognitivist theory of emotion, it is not the only important criticism of this theory. It has problems explaining recalcitrant emotions and emotional motivation, as previously noted, but also with attributing emotions to infants and animals, and with unconscious emotions.

What it is for an emotion to be fitting then is for it to appraise its object correctly. Whether that is so in any given case is an evaluative question about which people can differ – for instance, when they disagree about whether riding a bicycle without a helmet merits fear. But such differences on evaluative questions constitute real disagreement only insofar as there is some shared way in which their fear appraises things. It seems clear that this is true of many natural emotions, and it is possible to find a way of expressing that appraisal in language that all parties to such a dispute can accept.

2 The Motivational Theory of Guilt as a Natural Emotion

Moral philosophers tend to suppose that there is a sharp distinction between states such as anger and fear, which they typically grant to be psychological kinds and continuous with states of beasts, and those sophisticated social emotions with which philosophical moral psychology tends to engage, such as guilt, regret, shame, envy, and jealousy. Paul Griffiths (1997) argues that the former class constitutes a kind that, following Paul Ekman, he calls *affect programs*; but that the latter, which he calls the *higher cognitive emotions*, differs so drastically that the two classes do not belong to any common kind.

Subsequent critics have noted that Griffiths's influential argument for this popular distinction is hasty. His treatment of the affect programs understates the variety and complexity of states such as fear and disgust, which, at least in humans, are neither as systematically encapsulated from higher cognition nor as stereotypical in their behavioral output as he initially suggested.[10] And his focus on the differences between affect programs and higher cognitive emotions, such as the presence of clear biological markers and the automaticity of some of their symptoms, leads him to overlook motivational similarities that cut across this distinction. Even if some instances of fear, anger, and disgust form a biological kind as affect programs, there might also be a broader psychological kind that includes other instances of those emotions as well as guilt, jealousy, and the like.[11] Indeed, Griffiths seems open to this possibility in more recent

[10] Roberts (2003) makes this point among others against Griffiths's disunity argument.

[11] Prinz (2004) and Deonna and Teroni (2012b) press this point as well. Both also note that the category Griffiths calls "irruptive motivations" appears to include both affect programs and the examples of higher cognitive emotions mentioned earlier. We agree entirely on these points and develop them further, in what follows, by illustrating the explanatory power of the motivational theory in the case of guilt.

work (see Scarantino & Griffiths, 2011). Although some contemporary psychologists are skeptical that *any* emotions are natural kinds, we find their standards for such claims overly demanding and doubt some details of their arguments.[12]

Guilt is a good example of an emotion that recruits sophisticated cognitive faculties and lacks some of the physiological symptoms of bodily preparation for action characteristic of fear and anger, but which exhibits the peculiar motivational features distinctive of natural emotions. Guilt typically arises in response to voluntary action of the agent that gives others grounds for anger. Two familiar examples are personal betrayals (which give a specific person such grounds) and breaches of moral rules (which give them to all). Bouts of guilt display the *control precedence* characteristic of emotional motivation: they prioritize the emotional goal in attention and motivation. And they issue in actions such as confession, apology, and other direct and urgent effort to make amends, as well as in self-castigation – especially when restitution is impossible. Guilt is characteristically satisfied by indications that the injured party has accepted the apology, and that relations have been restored to something like the status quo ante. Hence, the goal of guilt seems best described as the *reparation* of some damaged relationship, either with a specific person or with the community at large.

Guilt exhibits the peculiarities characteristic of emotional motivation, despite its cognitive complexity. It can issue in acting without thinking, when it motivates overly direct and urgent means to meet its goal, or when it leads the guilt-ridden agent to sacrifice ends that are more important by his own lights. Actions performed in the throes of guilt are often insensitive to these other ends, and to some of the agent's information about how best to achieve the goal of reparation. Thus, people attempting to get away with wrongdoing can be undone by their guilt when it prioritizes the goal of reparation in ways they do not endorse on reflection. And even those who endorse reparation as their overriding goal can be prompted, by their guilt, to poor means of achieving it such as overapologizing, confessing too often or at too great length, and performing acts of contrition that predictably serve to discomfit the victim rather than repair the relationship. A hallmark of the emotions is their prioritization of a generic goal and narrowed attentional focus – that is, control precedence – and their

[12] An especially influential skeptic is Lisa Feldman Barrett (2017a). Her recent exchange with Ralph Adolphs illustrates some of the controversies within neuroscience (Adolphs, 2017a, 2017b; Barrett 2017b, 2017c). We address these issues at some length elsewhere (D'Arms and Jacobson, forthcoming) but will not pursue them here.

prejudice in favor of direct and urgent means to satisfy that goal. These similarities in what and how emotions motivate are common between so-called affect programs and some higher cognitive emotions, and this makes us skeptical about putting too much weight on that distinction.

Guilt is also susceptible to stable recalcitrance, in that you can feel strongly driven to apologize or make reparations for something you did, or even something that merely happened to you, despite your considered judgment that your guilt is unfitting. This can happen in cases of survivor guilt, for instance, when someone is convinced that he has done nothing wrong and yet continues to feel guilty. Recalcitrance is further evidence that guilt is a discrete source of motivation, despite its complexity, which can persist at odds with the agent's considered judgment. We think it plausible that guilt is an adaptation, which is part of normal human nature because it enabled our ancestors to respond to their own transgressions in ways that helped them maintain better relationships with others; but that claim is not essential to the theory or this chapter.

It thus appears that guilt, like fear and anger, is a distinctive kind of affect-laden motivational system that has a characteristic goal (of reparation) and motivates a distinctive way of pursuing that goal. That is, bouts of guilt prioritize the goal and direct cognitive resources toward its direct and urgent satisfaction, potentially at the cost of attending to its relative importance and whether the actions it urges are the best means of meeting its own goal. Its nature is a matter for empirical investigation, not for specification by conceptual analysis. It is not even partly constituted by a particular judgment or thought that can be given in response-independent terms, because the appraisal is an interpretation of the emotion as a whole, including its motivational aspects. The terms in which the gloss is given must therefore be understood in light of the emotion's goal, such that it appraises its object – in this case, one's own action – as giving one reason to act in reparation. How then should one interpret the generic appraisal of guilt, so as to understand the conditions under which it is fitting?

The procedure we previously outlined starts from an empirical characterization of guilt. While philosophers most often focus on guilt as a response to moral transgression, its paradigmatic elicitors actually fall into two broad kinds: not only actions involving moral violations, such as theft and murder, but also actions that constitute some sort of transgression against a personal relationship, like disappointing a loved one.[13] Its typical phenomenology involves feeling bad about what one has done, specifically

[13] Tangney and Dearing (2002) describe studies that support these commonsense observations.

for those it hurt, and the desire to express this feeling. In short, guilt is experienced as a felt desire to make amends. As noted, guilt motivates apology, confession, and efforts to compensate where possible; and it is most likely to be satisfied by sincere forgiveness or other signs that the relationship has been repaired. In light of these features, we suggest that someone in a bout of guilt can be interpreted roughly to appraise himself as having engaged either in some sort of *wrongdoing* or in a *personal betrayal*.[14]

If we are right that someone who feels guilty about something can be understood to take it as a wrongdoing or a personal betrayal, then those familiar with guilt should find this gloss plausible and agree that it sets the terms for assessing when guilt is fitting. But the gloss remains rough because of the point noted earlier: the way that you take something in the throes of an emotion is shaped by the character of the emotion itself. Hence, the relevant terms must be allowed enough semantic slack to accommodate the response-dependent evaluation they seek to articulate. In this case, they must be understood in a way that accommodates excuses. If someone was coerced into stealing in a way that you think renders guilt unfitting, for instance, then you think he has not really acted wrongly in the sense of that term that captures guilt's appraisal. One could say instead that guilt appraises what one has done as an *unexcused* wrongdoing or betrayal. Though this addition might avert certain misunderstandings, it creates others, since not everything that counts as an excuse in ordinary language, law, or social custom renders guilt unfitting. Instead, we will retain the simpler version, with the caveat that it is (inevitably) a rough-and-ready characterization of a response-dependent appraisal.

We consider some implications of this gloss in the final section. What is crucial for present purposes is that our account of guilt satisfies the success conditions previously given. We have argued that guilt satisfies the priority condition, because the motivational theory does not make it require any response-independent evaluative thought. Our gloss of its appraisal is not a constituent of the emotion but an articulation of what guilt concerns in light of its nature, which enables an account of when it is fitting. It is a further question whether guilt satisfies the rational assessment condition on an account of the blameworthy. This a matter of whether its nature

[14] Our account of guilt's appraisal is unconventional, and hence controversial, because it makes room for the possibility that guilt can be fitting over actions that are not morally wrong – and perhaps even obligatory. This will be the case for betrayals of an intimate for overriding impersonal reasons. We develop a case of this sort in D'Arms and Jacobson (1994) as part of an argument against Gibbard's (1990) neo-sentimentalist account of the blameworthy. In these cases, arguably *someone* has reason to blame you, namely the person whom you betrayed; but others have no such reason.

is such that it is appropriate, in the relevant respect, to feel guilty over just one's blameworthy actions. In order to assess this, we must offer an account of the relevant sense of appropriateness.

3 Appropriate Guilt: Fittingness, Not Desert

According to the Strawsonian dictum (*), for A to be blameworthy for *x* is for it to be appropriate for A to feel guilt for *x*. Clearly, the term "appropriate" is normative, in contrast to a dispositional view on which for A to be blameworthy for *x* is for A to be prone to guilt over *x*. Although there are more plausible forms of dispositionalism, we find them all inadequate. But "appropriate" is vague, and there are various ways to flesh it out that have disparate implications. If an attitude is said to be appropriate just in case it is optimal, for instance, then the dictum would be open to familiar counterexamples involving evil demons and eccentric millionaires who create incentives for having the attitude. Wallace (1994) has proposed that responsibility should be understood in terms of the *fairness* of the blaming emotions, but we agree with Carlsson (2017, p. 19) that certain considerations relevant to the fairness of blaming – like whether others have been blamed for similar actions – do not bear on whether an action is blameworthy (see also Vargas, 2004).[15] We will focus on what seem to us the two most promising ways to understand appropriateness in (*): as the claim that guilt is *fitting* and that it is *deserved*.

The cognitivist theory of emotion has a seemingly straightforward account of fittingness as the truth of an emotion's constitutive thought. We consider this a specious advantage of the theory, since cognitivists dispute exactly what is the constitutive thought of an emotion type, and their method affords them no good way to resolve such dispute. However that may be, we have now shown that the motivational theory of emotion can hold similarly that an emotion is fitting when its appraisal is correct. These appraisals are to be understood not as constitutive thoughts necessary for having the emotion, but as an overtly response-dependent interpretation of the emotional syndrome as a whole.[16] This allows us to hold that (*) should be given in terms of fittingness. Carlsson (2017) argues, to the contrary, that the relevant notion of appropriateness is that of desert.

[15] It may be that Wallace's view ultimately does not differ from that of Carlsson and Vargas, as Rosen (2015, p. 70) suggests, at least when it comes to guilt and self-blame.

[16] This means that both theories of emotion are compatible with what Rosen (2015) calls the *alethic* view. Rosen's terminology differs from ours, since he contrasts the alethic view of appropriateness with a view of appropriateness as fittingness. But Rosen uses "fittingness" for a different notion than that of correctness, specifically as a primitive normative notion.

The claim that someone deserves to feel guilt over what he has done is a moral assessment, but Carlsson's proposal is not simply the retributivist intuition that the blameworthy deserve to feel guilty. Although that claim is controversial, we find it plausible.[17] Rather, the question at hand concerns what it is to be blameworthy; specifically, whether the Strawsonian account is better off understanding appropriateness as fittingness or desert.

There is no tension involved in holding guilt to be both fitting and deserved in some circumstance, or even in thinking that guilt is deserved whenever it is fitting. Nevertheless, these are distinct notions that figure differently in an account of blameworthiness. We will argue that fittingness is the best way to understand appropriateness within a Strawsonian framework, on two grounds. First, this approach treats blameworthiness analogously to other sentimental values such as the funny, the shameful, and the disgusting. Second, it is plausible that guilt is deserved only when, and because, it is fitting.

We have argued elsewhere that a number of values are best understood in sentimentalist terms, as the appropriate object of some associated attitude (see D'Arms & Jacobson, forthcoming). Call these the *sentimental values*. What it is for something to be dangerous is for it to merit fear; to be shameful is to merit shame; and to be admirable is to be a fitting object of admiration. In each case, the relevant form of appropriateness is fittingness or merit, understood as a notational variant: for an emotion F(*x*) to be fitting is for its object *x* to merit F. When something endangers a person sufficiently, it is fitting for her to fear it; however, she typically will not *deserve* to feel afraid, nor does the dangerous thing deserve her fear. If someone is especially graceful or beautiful, she may not deserve to feel proud of this trait, but pride is fitting simply because it reflects well on her. Similarly, for the other sentimental values, the relationship between sentiment and value is that of fittingness. Why should blameworthiness be different?[18] The proposal to understand appropriateness as fittingness gives a consistent theoretical treatment to an array of response-dependent values.

[17] It has been defended recently by Randolph Clarke (2016) and criticized by Dana Nelkin (2019). And, of course, it will be rejected by those who are skeptical of desert in general, such as Pereboom (2014).

[18] One might object that blameworthiness is different because to be blameworthy for some action is obviously tied up with being responsible for it, and being responsible for something bad makes you deserve your feelings of guilt. In contrast, one can have a shameful trait for which one is in no way responsible; while shame for such traits is fitting, one does not *deserve* to be ashamed. We accept this difference, but we think it is to be explained by the differences between guilt and shame that affect when they are fitting, not by a difference in the kind of emotional appropriateness involved in being blameworthy for something versus having a shameful trait.

Compare how each proposal explains why guilt is appropriate in those cases where it seems to be (such as for a robber), and inappropriate in other cases (as with survivor guilt). Our answer is that guilt over robbing someone is fitting – and therefore appropriate in the relevant sense – because it gets matters right. The robber's guilt appraises his action as wrong, and it is. The survivor's guilt appraises her survival in the same way, as a wrongdoing or betrayal, but it is not. The reason that survivor guilt is inappropriate is that she did nothing wrong and betrayed no one. The desert approach holds that guilt over having robbed someone is appropriate because the robber deserves to suffer for it – specifically to feel guilt over what he did – whereas the survivor did nothing to deserve that.

The general retributivist claim that the blameworthy deserve to suffer for what they have done is not tantamount to the claim that they deserve to feel guilt in particular. The thought that someone deserves to suffer does not differentiate between the suffering of a toothache and the suffering of feelings of compunction; all that matters for this basic retributivist intuition is that the suffering is proportionate to the degree of blameworthiness. A Strawsonian account of blameworthiness in terms of deserved guilt, on the other hand, is an attempt to explain blameworthiness in terms of that specific attitude. It therefore must explain why *guilt* is distinctively appropriate.

The answer must have to do with the nature of guilt, by virtue of which blameworthiness can be understood in terms of desert of this specific emotion. Carlsson characterizes guilt as a combination of painful affect and propositional content somehow held in mind – whether as a belief, thought, or seeming. Although the details are left somewhat vague, he proposes that guilt's propositional content is that the agent displayed an objectionable quality of will (2017, pp. 101–102). This suggests that what makes guilt the deserved form of suffering, which enables it to play a role in the analysis of blameworthiness, is that it involves being pained specifically at the thought that one displayed an objectionable quality of will.

But when does someone deserve this form of suffering in particular? The obvious answer is that one deserves this form of suffering just when that thought is correct. You do not deserve to suffer guilt in particular, as opposed to any other form of pain, unless the thought distinctive of guilt is true. If that is right, however, then the judgment that guilt is deserved is parasitic on the judgment that it is fitting. The point at hand does not depend on whether guilt takes its object as a wrongdoing or betrayal, simply as a wrongdoing, as an act of ill will, as all of those things, or some other thing. Nor does it matter if this content is to be found in a belief,

thought, seeming, or appraisal. The point is simply that in order for guilt in particular to be deserved on the basis of what one has done and why one did it – as it must be to play the relevant role in an account of blameworthiness – its content must be correct. And that entails that in order for guilt to be deserved, it must be fitting.

Even so, defenders of the desert proposal might object that our argument shows only that it is necessary for guilt to be fitting in order for it to be deserved. It would still be possible for guilt to be fitting over some action but not deserved. According to this objection, such an action would not be blameworthy. It is unclear to us how best to make this argument, but perhaps a rationale can be found in Carlsson's claim that desert, unlike fittingness, accounts for the idea that in order for someone to be blameworthy, it must be noninstrumentally good that she suffers.[19] The idea that it is good for a wrongdoer to suffer is controversial, of course, even among those who accept that people can be blameworthy – not just that blame can be beneficial. It seems to us an advantage of fittingness that it can accommodate those who favor a Strawsonian account but are skeptical about the claim that it is noninstrumentally good for wrongdoers to suffer. And even those who grant the axiological claim can hope that the fittingness of guilt explains why that state in particular constitutes the sort of suffering that wrongdoers deserve.[20]

We conclude that fittingness best captures the appropriateness of guilt invoked in (*). Hence, a Strawsonian theory of blameworthiness that focuses only on self-blame should hold that for A to be blameworthy for *x* is for it to be fitting for A to feel guilty for *x*. We have offered an account of guilt and its fittingness that allows the theory to avoid the challenges of circularity and otioseness. Even if a full account of blameworthiness ought to appeal to other reactive attitudes as well, we think it plausible that the fittingness of guilt provides one central human mooring on which our concepts of blameworthiness and responsibility depend.

[19] The claim that deserved guilt is noninstrumentally good is defended by Clarke (2013), and Carlsson cites his arguments for this claim approvingly. In our view, some of those arguments apply equally to fittingness. But Clarke is not arguing for an account of blameworthiness in terms of appropriate guilt, nor for a preference for desert over fit within such an account.

[20] Judgments of fittingness are normative. They involve thinking that there are reasons to have the emotion and, moreover, reasons to act in some way that is relevant to satisfying its goal. This suffices to give at least partial endorsement to feelings and actions that would constitute costs to the blameworthy party. We think this captures what is correct in the elusive intuition, which – like all claims about noninstrumental goodness – is difficult to argue for directly.