Plasticity and perfection:
Emotions and the moral domain

Gopal Sreenivasan
gopal.sreenivasan@duke.edu

Abstract According to de Sousa's famous 'new biological hypothesis 2,' '[e]motions are species of determinate patterns of salience among objects of attention, lines of inquiry, and inferential strategies.' Among other things, this turns out to mean that a particular emotion functions both to recognize and to focus the agent's attention on instances of its characteristic object, as well as to motivate a characteristic suite of responses to that object (in and) by the agent. Quoting de Sousa further, we can say that emotions thereby effect a double 'control of salience,' once on their input side and again on their output side. In this paper, I explore the question of how much developmental plasticity emotions exhibit in their performance of this double control. To this end, I describe two different models that can be used to demonstrate the plasticity of an emotion, which I call the 'fixed culture' model and the 'variable culture' model, respectively. My discussion is limited to the examples of fear and sympathy (one a 'basic' emotion, the other not). The relevance of my conclusions about emotional plasticity is mediated by the conjecture -- stated here, but not defended -- that an emotion's double control of salience is epistemically useful, both in general and in the moral domain specifically. Most importantly, the plasticity of emotion allows for its epistemic contributions to be improved, and to that extent, perfected.

As is perhaps fitting in our ecological age, one of my favourite stories from de Sousa's marvellous oeuvre is actually a recycled story. In *The Rationality of Emotion* (1987), Ronnie borrows a story from Dennett (1987) to introduce one of his central claims about emotion. The protagonist of this story is a robot on whose wagon there happens to be a bomb. Now the robot not only knows where the bomb is, but also that it is about to go off. However, and unfortunately for the robot, its programming instructions are inadequate to enable it to recognise the relevance of this knowledge. So the poor thing ends up getting blown to bits. Some of the humour in the story derives from the repeated failures of the engineers to avert this disaster by re-programming their robot (at one point, e.g., they instruct the robot to ignore irrelevant implications of what it knows). The moral suggested by the tale, which de Sousa boldly affirms, is that pure reason is insufficient to determine what is relevant.

In de Sousa's hands, this moral is merely a prelude to a brilliant conjecture about the function of emotion, according to which emotions function to solve the problem of determining

what is relevant.[1] Alternatively, they function to make up for the 'insufficiencies of reason,' something they accomplish 'by controlling salience,' as de Sousa memorably puts it (201).[2] To quote his elaboration of this suggestion:

> [for] a variable but always limited time, an emotion limits the range of information that the organism will take into account, the inferences actually drawn from a potential infinity, and the set of live options among which it will choose (195).

The case of fear provides a pithy illustration of what de Sousa has in mind. If the robot in his story had been equipped with fear, this fear would very likely have been triggered by the robot's knowledge of the state and location of the bomb. Among the various consequences of its thus being occurrently afraid, two consequences fit de Sousa's description quite precisely: the robot's attention would then have been focused on the bomb (i.e., the danger at hand) and its set of live response options would likewise have been limited to 'fight or flight.'[3] '[Actively] ignoring thousands of irrelevant implications' would certainly not have remained as one of its live options. In this way, emotion is able to outdo the engineers.

For my own part, I was so taken with de Sousa's conjecture about emotion that I went to town with it. That is to say, I used it as the foundation for an account of the moral psychology of virtue, on which I (eventually) wrote a whole book.[4] Naturally, I am not proposing to rehearse any of that account here. After all, our aim is to celebrate Ronnie. What I should like to examine, rather, is the suitability of emotion to contribute to moral reasoning more generally—where by

---

[1] Ronnie calls this his 'new biological hypothesis.'

[2] Bare page references are to de Sousa (1987).

[3] Of course, fleeing the bomb was impossible under the circumstances. In that case, the robot's fear would presumably have focused its response on some version of 'fight' [-ing the bomb]. For a nice discussion of the inadequacy of 'fight or flight' as a complete menu of fear responses, see Tappolet (2016: 53-56).

[4] *Emotion and virtue* (2020). Hereinafter, *EV*.

'contribute,' I mean the same epistemic contributions that feature in de Sousa's idea about controlling salience. By contrast to my previous undertaking, however, I shall not pay any special attention to virtue, despite the fact that virtue plainly involves moral reasoning.

In fact, I shall not even describe moral reasoning very closely. I shall begin by extracting a simple and fairly obvious necessary condition that emotions would have to satisfy, in order for any epistemic contribution they may make to be eligible to count as part of good moral reasoning. My principal claim will be that emotions do satisfy this necessary condition. On this basis, I shall suggest that Ronnie's lesson about emotions is robustly suitable to being applied in the moral domain. To that extent, it is all the more valuable.

Let me say a little bit more up front about the framework that scaffolds the more detailed discussion to follow. This capsule preview will furthermore serve to illuminate my title. The necessary condition to which I referred is that emotions be (morally) *perfectable* in certain ways, where their perfectability in turn requires them to be developmentally plastic. The main burden of my argument will be to demonstrate that emotions are indeed developmentally plastic. Here I shall concentrate on two examples, fear and sympathy, which usefully straddle the division between basic and non-basic emotions. Along the way, I shall also distinguish two rather different models of how developmental plasticity can be exhibited, which I shall call the 'fixed culture' model and the 'variable culture' model. These models are complements rather than rivals and my discussion of the plasticity of emotion trades in both models.

§1

To mark a space within which a need for this or that emotion to be morally perfectable can be clearly grasped, let me distinguish various senses in which an emotional reaction can be considered

'appropriate.' I shall start with what are ordinarily considered as emotional reactions, which puts us on the territory inhabited by the second consequence privileged by de Sousa's idea that emotions control salience. In the case of fear, this was illustrated by the fact that fear motivates a specific *response to* danger, namely, fighting it or fleeing it. Subsequently, I shall extend this apparatus (and the claims I make with it) to the first consequence privileged by his idea—in the case of fear, that it focuses the subject's *attention on danger* in its environment—and thereafter I shall synthesise and reprise my whole position for the case of sympathy.

Now the most obvious sense in which fleeing some particular danger can be seen as an 'appropriate' response to that danger is that flight preserves the organism's life. It is therefore appropriate in the sense of being functional or fitness-enhancing. From an evolutionary perspective, this functionality of the responses to (apparent) danger that fear characteristically motivates goes hand in hand with the idea that fear has been selected for. In addition to this obvious sense of appropriateness, I should like to distinguish two further senses.

On the one hand, there is what we might call 'emotional' appropriateness, for want of a better term. As Ronnie (among other people) has emphasised, emotions have a narrative structure. Thus, the components or elements of a typical emotional episode can be perspicuously represented in terms of a script, where this script is not only specific to the emotion in question, but also tells some kind of story. In this sense, fleeing is an appropriate response to danger insofar as it is a fitting or familiar ending to a danger story (as is fighting).

Unlike functional appropriateness, emotional appropriateness does not depend on the subject of the emotion's having *true* beliefs about the object of her emotion. Fleeing witches or tilting at windmills can make perfect narrative sense. By contrast, fleeing an apparent danger is not functional on the occasion unless the apparent danger is also real. More generally, having fear

as a trait is not functional (or adaptive) unless the identifications of apparent danger 'written in' to the trait are roughly reliable as identifications of real danger (at least in the period of adaptation).

On the other hand, there is *moral* appropriateness. Moral appropriateness is logically independent of functional appropriateness, with the former being neither necessary nor sufficient for the latter.[5] Fleeing some danger may well be life preserving, and therefore functionally appropriate (not to mention, the action the subject is highly motivated to undertake). But for all that, fleeing may still be cowardly, and therefore morally inappropriate.[6] Likewise, fighting some danger—an enemy combatant, say—may be morally required despite the fact that it leads (more or less directly) to the subject's death. When this outcome is known in advance to be overwhelmingly likely (i.e., known to the subject, too), fighting cannot be functional for the individual—not on the occasion, anyhow. However, this twist is perfectly consistent with fighting's remaining morally required.

Suppose we accept that fear plays a useful role in guiding ordinary individuals to select 'fight or flight'—rather than any of the alternatives, including doing nothing—as their practical response to a dangerous situation, and so makes an epistemic contribution to their generic practical reasoning. Under what conditions does this same contribution count, furthermore, as a contribution to the individual's moral reasoning? At a minimum, it would seem, the selection an individual's fear makes among her available responses to a given danger *only* counts as a contribution to her moral reasoning *when* the response thereby selected is a morally appropriate response. Insofar as the response fear selects is morally inappropriate, the operation of fear inhibits

[5] For a similar claim, see D'Arms and Jacobson (2000), who dub various breaches of this independence the 'moralistic fallacy.'

[6] I discuss the relations between the characteristic action tendency in fear and judgements of courage and cowardice in *EV*, ch. 9.

the individual's moral reasoning (instead of contributing to its success). Yet nothing in the nature of fear prevents it from selecting morally inappropriate responses.

Of course, on any given dangerous occasion, an individual might well be able to overcome such inhibitions to her moral reasoning as fear may cause. For example, she may be able to control or resist her occurrent fear. On some conceptions, including the common sense one, this possibility illustrates the primary challenge in courage. In principle, however, there is an alternative to this occurrent control strategy, namely, modifying the responses that an individual's fear disposes her to select in the first place (i.e. modifying her underlying fear trait itself). More specifically, what the alternative—or perhaps it is only a supplement—contemplates is to modify an individual's fear trait by *subtracting* some number of the morally *in*appropriate fight or flight responses her fear will dispose her to select. The greater the number or variety of morally inappropriate responses that are subtracted, the less morally imperfect this individual's fear trait will be. Likewise, the more reliable its epistemic contributions to her moral reasoning will be.

Here we reach the question of how far fear is actually amenable to such a developmental control strategy, as we might call it. While it is coherent in principle, how far can this strategy be realised in reality? By this I mean, to what extent does *fear* lend itself to such modifications (or admit of them)? I do not mean to ask how far the social or mundane obstacles to taking advantage of any plasticity fear may have can be overcome. What interests me, for present purposes, is the plasticity of emotion per se, rather than the prospects for social or educational reform. I shall begin by arguing that fear exhibits significant plasticity with respect to the responses to danger it sets in motion when triggered.

Since my argument on this score follows a specific model for exhibiting an emotion's developmental plasticity, I should first introduce that model. According to what I call the *variable*

culture model of plasticity, the crucial question concerns the extent of cross-cultural variation in the emotional phenomena of interest. Our initial focus is specifically on emotional responses. In that case, the model maintains that the more cross-cultural variation there is in the responses characterising occurrent fear (say), the more plastic fear responses are shown to be. Take facial expression, for example, which is one of the most studied components of an emotional response. Here the variable culture model holds that the greater the variation across cultures in the facial expression of individuals experiencing fear, the more plasticity the facial expression of fear has.

Let me spell the basic thought underlying this model out. Suppose we have two cultures, C1 and C2. In C1, fear has one distinctive facial expression (F1), whereas in C2 it has a different distinctive expression (F2). This supplies the cross-cultural variation which constitutes the model's starting point. Now consider an infant orphan in C1, Omni. Ordinarily, the later adult Omni will have a fear disposition that includes the expression of F1 when triggered. However, if Omni were to be adopted in infancy by a family in *C2*, that is not how she would develop.[7] Rather, the adult Omni would then have a fear disposition that instead includes the expression of F2 when triggered. Thus, the infant Omni's fear disposition is 'developmentally plastic,' since it can develop along either of these two pathways. Moreover, the same set of facts exhibits the possibility of *subtracting* the expression-of-F1 component from the infant Omni's fear disposition—subtracting it, that is, relative to that emotion's default development in C1—by transporting her to C2 (and having her grow up there instead).[8]

---

[7] I begin with the idea of cross-cultural adoption. Later we shall encounter other (more widely practised) modalities that have the same effect.

[8] Although I emphasise subtraction from an emotion's default development in the text, the changes enabled by variable culture plasticity are symmetrical. Thus, the same transportation of the infant Omni could equally well be seen as leading to the *addition* of an expression-of-F2 component to her fear disposition. Or if there are three cultures and an option of F3 in some C3,

My variable culture argument that fear responses exhibit significant developmental plasticity is simply to observe that the variables in the schema outlined in the previous paragraph can be replaced with specific descriptions and, provided that one's selection has been judicious, the resulting claims will turn out to be empirically well confirmed. A very straightforward illustration can be found in the work of Paul Ekman and his colleagues (see, e.g., Ekman 2007).

Ekman defends a version of what is often called the 'affect program' theory of emotions. The central idea of this theory, to quote Paul Griffiths, is that 'emotional responses are complex, coordinated, and automated' (1997: 77). Among the response elements coordinated by a given affect program are changes in facial expression (i.e., the very example we used earlier). According to Ekman and his colleagues, there is a specific and recognisable facial expression for fear, as well as for anger, sadness, disgust, and happiness. However, they importantly distinguish between two very different kinds of facial expression, natural expressions and social expressions.

This distinction turns on Ekman's notion of 'display rules' for emotion. Display rules regulate when and how (and to whom) a given emotion can be displayed (e.g., expressed in one's face).[9] For example, they may 'dictate that we diminish, exaggerate, hide completely, or mask the expression of emotion we are feeling' (Ekman 2007: 4). *Social* facial expressions of emotion are facial expressions that are modulated by some display rule, whereas *natural* facial expressions are not modulated by any display rule. Since a display rule may always permit the open, unmodulated

---

we could say there is room to choose whether to add F2 or F3, in place of the F1 that will be subtracted by Omni's adoption out of C1.

[9] Display rules can be understood as operationalising a fourth sense in which an emotional response can be 'appropriate.' They specify the conditions under which a particular emotional response is *socially acceptable* to display. Display rules thus regiment a form of etiquette.

facial expression of some emotion, a particular facial configuration can count as both a natural and a social facial expression.

When Ekman claims, then, that there is a specific and recognisable facial expression for each of the affect program emotions, he is making a claim about their natural facial expression. Indeed, for natural expressions, Ekman and his colleagues go further still, claiming that the emotion-specific facial expressions they identify are *pan-cultural*. Summarising a large body of empirical research, both their own and that of others, Ekman, Friesen, and Ellsworth (1982: 141) report that the

> same emotions were judged for the same facial behaviors by observers from different cultures in experiments that had many different stimuli of many different stimulus persons and many different groups of observers from 14 cultures or nations. Similar results were obtained with visually isolated, preliterate, New Guinea observers.

By contrast, Ekman does not claim that social facial expressions of affect program emotions are pan-cultural. Rather, on his account, the display rules for the facial expression of a given emotion will vary significantly between cultures. It is therefore his evidence for this proposition that illustrates the variable culture argument for the plasticity of the social facial expression of any number of affect program emotions.

To that end, let us review one of their famous experiments. Ekman and colleagues arranged for college students in Berkeley and Tokyo to watch a pair of short films—25 in each place. One was a neutral film, while the other was stress inducing. On the first iteration, each student watched the films alone, without knowing that he was being videotaped. In both groups, a marked difference was found between the facial expressions shown during the two films. During the stress film, the students showed many more expressions scored as surprise, disgust, sadness, and anger (compared to during the neutral film). Moreover, the repertoire of facial expressions was very

similar between the two cultures. Ekman and colleagues interpret this as confirming their hypothesis of cross-cultural uniformity in the natural facial expression of affect program emotions.

But the second iteration in the experiment is the decisive one for our purposes. Each student watched the stress film again, now in the company of a scientist from his own culture. This intervention resulted in a dramatic difference in the facial expressions exhibited as between the two groups of students (under otherwise identical eliciting conditions as before). For this time, the Japanese faces expressed far fewer negative emotions than did the American faces. Instead, they displayed many more polite smiles. Evidently, the Japanese students (but not the Americans) had learned to mask facial expressions of negative emotion in public with polite smiling. In other words, the Japanese and American students had succesfully internalised very different display rules for negative emotion.

Hence, returning to the schematic version of the variable culture argument, we can let C1 and C2 be the United States and Japan, respectively. Similarly, F1 can be the natural facial expression that Ekman identifies as the pan-cultural expression of fear and F2 can be a polite smile.[10] Let Omni be an American infant orphan. Her fear disposition is plastic, then, at least to the extent that it can develop in one of two rather different ways. Either there will later be no difference in her facial expression when she is afraid as between being afraid alone or in public (if she grows up in America) or her natural fearful expression will later be reflexively replaced with a polite smile when she is afraid in public (if she grows up in Japan). Alternatively, Omni's fear disposition is significantly plastic insofar as it lends itself to being reflexively controlled (in its

[10] In the solo viewings of the stress film, the emotion most commonly expressed was actually disgust. But the relevant display rules apply to the expression of negative emotion generally.

facial expression) when conditions are such that its default manifestation is 'socially inappropriate,' where these conditions are given by the display rules in force in Japan.

§2

Let us return to Ronnie's recycled robot tale. Recall that adding fear to the robot's equipment, so Ronnie can be taken to suggest, would plausibly have improved the robot's epistemic situation in two quite different ways. On the one hand, it would have resulted in the robot's attention having been focused on the danger at hand; and on the other hand, it would have resulted in the robot's live response options having been limited to 'fight or flight.' So far I have concentrated on the second kind of epistemic contribution that fear makes (or can make), but now I should like to attend some to the first.

Part of the way emotions control and focus the subject's attention results from a division of labour among specific emotions. Each specific emotion corresponds to a distinct domain of relevance, which at a very abstract level can be identified in terms of what is sometimes called the emotion's 'characteristic object' (compare, e.g., Ronnie's ch. 5). Thus, fear focuses the subject's attention on *danger*, as I have said, while anger focuses it on *insults or frustration*, disgust on *noxiousness*, and so on. When a subject's attention is focused on danger, rather than on insult or on something irrelevant, this can often be explained at least in part by the fact that it is her fear that has been triggered (rather than her anger or no emotion at all).

In addition to being distinguished from each other abstractly, in relation to their respective characteristic objects, specific emotions can also be distinguished *concretely*, by means of the different particulars or conditions in the world that trigger or elicit their occurrence. We can refer to these particulars as the 'triggers' or 'eliciting conditions' for a given emotion. For example, in

our imagined continuation of the robot tale, the trigger for the robot's occurrent fear is the bomb on its wagon. Standardly, when a given emotion is triggered, it does not merely focus the subject's attention on the abstract category corresponding to that emotion's characteristic object, but rather focuses her attention more specifically on the very particular that triggered its occurrence (e.g., on the bomb).[11] In this way, the occurrent emotion also categorises and evaluates the trigger in question for the subject (e.g., as dangerous). The set of triggers or eliciting conditions that partly define a subject's fear disposition, say, can therefore also be understood as the operational interpretation her fear gives to the extension of 'danger.'

Once we have distinguished the abstract category that serves as a given emotion's characteristic object from the various concrete particulars belonging to an interpretation of that category's extension, we are in a position to ask whether—and in what sense—it is 'appropriate' to treat a given particular as an instance of this or that characteristic object. Is it appropriate, for example, to treat a bomb in one's vicinity as a danger? In what sense of appropriateness? To some extent, we could always reproduce our previous distinctions among senses of appropriateness here and enquire accordingly. For example, if the bomb is live and set to go off, it is presumably functionally appropriate to treat it as a danger.[12] It is an interesting question whether there is room

[11] Of course, there is plenty of room for things to go wrong, so not every emotional episode will conform to this 'standard.' For example, in cases of emotional misattribution, the occurrent emotion focus the subject's attention on something other than the actual trigger for the episode. Furthermore, in standard and deviant cases alike, the trigger need not even exist, let alone be present in the subject's immediate environment. As everyone knows, emotions can easily be engaged in memory, imagination, or fiction.

[12] One might be tempted to say that it is functionally appropriate to treat some particular as a danger just in case it is correct to treat it as one. But this will depend on one's analysis of the correct extension of an emotion's characteristic object, which is controversial. On some analyses, the extension of an emotion's characteristic object cannot be fixed correctly without invoking some measure of response dependence. That is why some philosophers prefer to describe the

to wield 'moral appropriateness' as a sensible basis on which to evaluate the inclusion of this or that particular within the category of 'danger.'

Nevertheless, I shall not pursue that line of enquiry myself. For there is a much simpler path, albeit an indirect path, to bringing evaluations of moral appropriateness to bear on the relation between specific eliciting conditions for some emotion and that same emotion's characteristic object. This path begins from the fact that, taken together, the input and output sides of a given emotion serve to *pair* a specific set of responses *with* a specific set of eliciting conditions. That is plainly one of their effects, if not also one of their functions. As a result, when a specific eliciting condition operates as an instance of a given emotion's characteristic object, any agent who has that emotion is thereby primed to respond in certain specific ways to *that* condition. Returning to our running example, if the robot's fear program treats *bombs* as dangers, this means that the robot will (or is certainly very likely to) respond to a bomb *by* fighting or fleeing *it*. Details about the extension of an emotion's characteristic object therefore engage indirectly, but quite straightforwardly, with familiar evaluations of the moral appropriateness of the emotion's characteristic responses.

This is enough to secure a foothold for ambitions of moral perfection on the input side of emotion, and with them our question of developmental plasticity. In general terms, plasticity in relation to the eliciting conditions for some emotion is a matter of whether these conditions can be changed (or how far they can be), where change comprises both adding particular eliciting conditions to, and subtracting them from, an individual's emotional disposition.

---

characteristic object of fear (say) as 'the fearsome,' instead of as danger. For helpful discussion, see Tappolet (2016, ch. 3).

To illustrate the connection between the plasticity of an emotion's eliciting conditions and perfecting its epistemic contributions to the subject's moral reasoning (i.e., reducing their imperfections), let us consider fear in human beings rather than in robots. Within an evolutionary perspective on emotion (favoured, e.g., by affect program theories), it stands to reason that reliable eliciting conditions for fear include genuine physical dangers—or at least conditions that were genuine physical dangers in the period of evolutionary adaptation. Snakes, spiders, and heights are examples of such 'evolutionarily relevant fear stimuli' (Öhman 2010). However, on the very plausible assumption that some modern physical dangers (bombs, say) and some social dangers (censure or ostracism, say) are *also* genuine dangers, it would obviously be advantageous if novel conditions could be added as reliable eliciting conditions for fear.

More specifically, the addition of novel eliciting conditions to fear can reduce imperfections in its epistemic contributions to moral reasoning. Suppose, for example, that the scope of courage extends beyond 'physical courage' to comprise 'moral courage' as well. In that case, individuals whose fear is not (reliably) sensitive to *social* dangers are liable to make various moral mistakes about courage.[13] Certainly, their fear is unable to make the same epistemic contributions it can be relied upon to make in cases where the individual faces some (ancient) physical danger. At the same time, precisely this imperfection could be remedied if it were possible to add social dangers as novel eliciting conditions, i.e. if fear were plastic in that respect (too).

[13] Here I am assuming both that courage is at least partly a matter of the comparative evaluative balance between some goal the agent pursues and a danger in the face of which the agent pursues it; and that judging this balance correctly presupposes the agent's awareness of the danger. Thus, agents who lack awareness of some danger they face at least miss out on an occasion for courage and worse, may well act rashly. For further discussion, see my *EV*, ch. 9.

In the previous section, I argued that fear was developmentally plastic in relation to the responses it coordinates to danger. Here I shall argue that fear is also plastic in relation to its *eliciting conditions*. However, whereas my previous argument followed the variable culture model of plasticity, the present argument introduces and then follows a different model of plasticity, which I call the *fixed culture* model.

On this second model, the crucial question concerns the availability of some idiosyncratic course of development, by means of which one or more eliciting conditions can be added to the default set for a particular individual's emotion trait (or subtracted from it). The 'default set' are the eliciting conditions the individual's emotion trait would otherwise have (in the absence of a given intervention) and an 'idiosyncratic' of development is one that can be undergone by a minority of inhabitants in the target culture (and possibly, just one). Given some such course of development, it will be possible to add (say) new triggers to some individual's emotion trait *without* having to change the entire surrounding culture in which she develops.[14] Hence, 'fixed' culture.

The best example of such an idiosyncratic course of emotional development is classical fear conditioning. Its basic premiss is that, by simple and repeated pairing of a neutral stimulus (such as an auditory tone) with a naturally aversive stimulus (such as an electric shock), the subject will come to acquire the neutral stimulus as a learned trigger for the fearful response she is already prone to display to the naturally aversive stimulus (i.e., the 'unconditioned' stimulus). Subsequent to the conditioning, exposure to the neutral stimulus all by itself (i.e., to the 'conditioned' stimulus) will elicit a full-blown fear response from the subject. More specifically, the conditioned stimulus

[14] To change the entire surrounding culture, one has either to reform the individual's original culture wholesale or to transport her to a different culture prior to her development.

alone will then trigger the automatic appraisal mechanism (e.g., the focusing of attention) that is characteristic of fear (LeDoux 1996: 174-78; Phelps and LeDoux 2005).

For our purposes, the crucial feature of classical fear conditioning is that the conditioned stimulus can by any arbitrary antecedent. For instance, in many studies of fear learning in humans, it is a coloured square. A fortiori, a learned fear trigger can be any specific social danger that needs to be added to someone's fear trait. Since I cannot hope to improve on Joseph LeDoux's summary (1996: 143), I shall resort to quoting it:

> Fear conditioning opens up channels of evolutionarily shaped responsitivity to new environmental events, allowing novel stimuli that predict danger (like sounds made by an approaching predator or the place where a predator was seen) to gain control over tried-and-true ways of responding to danger. The danger predicted by these *learned trigger stimuli* can be real or imagined, concrete or abstract, allowing a great range of external (environmental) or internal (mental) conditions to serve as [conditioned stimuli].

Thus, given that the basic mechanism of fear conditioning is demonstrably effective—and, indeed, highly effective—the 'idiosyncratic course' of development that the fixed culture model of plasticity requires to be available plainly *is* available, at least in the case of fear. Moreover, since any given social danger falls within the scope of this mechanism's efficacy, the extent of fear's developmental plasticity in relation to its eliciting conditions is correspondingly wide. On this basis, the argument from fixed culture plasticity makes itself.

§3

So far I have advanced two arguments about the developmental plasticity of fear, one holding that fear is plastic on its input side and the other that it is plastic on its output side. In other words, at least at the outset of a given individual's development, the eliciting conditions for her fear disposition admit of being changed; and the responses that will be set in motion as part of any

episode of her fear admit of being changed, too. I also described two different models of how the plasticity of a given emotion may be demonstrated, the variable culture model and the fixed culture model. While I happened to use the fixed culture model in demonstrating the plasticity of fear's eliciting conditions and the variable culture model in demonstrating the plasticity of its responses, that mapping of models to tasks was arbitrary. As we shall see, it could easily have been reversed.

Now, in spelling each of these arguments out, I married the details of my specific application to fear to *affect program* theories of fear (e.g., to Ekman's theory). Like any developed theory of emotion, affect program theories are controversial and their merits can be debated. However, this particular tethering was a strategic choice, not a committed one. Let me explain. In their illuminating survey on the plasticity of emotion, Luc Faucher and Christine Tappolet (2008) distinguish three broad classes of theories of emotion according to how much plasticity the theory in question permits emotions to have. At one extreme, 'strongly determinist biological' theories permit the least plasticity, while at the other extreme 'social constructionist' theories permit more or less unlimited plasticity. Affect program theories are the paradigm of strongly determinist biological theories, which Faucher and Tappolet (2008) proceed to criticise for understating the extent to which emotions are plastic.[15] I take no position here on which class of theory gets the extent of the plasticity of emotions right. Nor do I need to.

My point, rather, is that affect program theories usefully provide a 'minimum index' of the plasticity of emotions. If this class of theory is mistaken, as well it might be, then emotions have even more plasticity than we have already seen (with affect program fear). However, if this class of theory credits emotions with *less* plasticity than any other theory does, as Faucher and Tappolet

[15] They favour an intermediate option, 'developmental systems' theories.

maintain, and yet the plasticity with which it credits them is nevertheless significant, then the worst case scenario is that emotions have significant plasticity. This, I claim, is precisely what we have seen in the case of fear, albeit with more structure and in greater detail. It is also the position I seek to establish.

Indeed, fear as conceived by affect program theories actually provides a minimum index of the plasticity of emotion along two quite different dimensions. The first dimension is theoretical. Along this dimension, *affect program* fear provides a minimum index of plasticity 'as compared to other *theories*' (i.e., as compared to the plasticity they permit fear to have), as I just explained. By contrast, the second dimension is emotional, or emotion-specific. Along this dimension, affect program *fear* provides a minimum index of plasticity 'as compared to other *emotions*.' From one perspective, fear really only stands in here for any 'basic' emotion, as compared with 'non-basic' emotions.[16] Provided one accepts this way of drawing that distinction, the point could be put equally well by saying that basic emotions provide a minimum index of plasticity as compared to non-basic emotions. Hence, non-basic emotions will be at least as plastic as basic emotions. What validates this inference is the proposition that non-basic emotions are *more cognitive* than basic emotions, where more cognitive emotions are expected to be more plastic because they are more culturally infused or less fully 'hard-wired.' Griffiths (1997) even calls these other emotions (i.e., those not listed in note 16), the 'higher cognitive' emotions.

On this basis, one is licensed to infer that sympathy, say, is at least as plastic as fear is; and this certainly seems to be a very reasonable inference. Still, in the remainder of this paper, I shall

[16] The perspective, that is, of affect program theories themselves, which conceive 'basic' emotions as the very emotions they privilege —fear, anger, sadness, disgust, happiness, and surprise.

reprise my arguments about the plasticity of emotion using sympathy as my example instead of fear. I have two aims in arguing for these conclusions explicitly, as opposed to reaching them via the reasonable inference. To begin with, relitigating the arguments for a non-basic or non-affect-program emotion (such as sympathy) serves to confirm or reinforce the reasonable inference, thereby increasing our confidence to rely on it in other cases. Furthermore, and perhaps more importantly, the connection between the epistemic contributions made by emotions and moral reasoning is more obvious in the case of sympathy than it is in the case of fear. Since my rationale for investigating the plasticity of emotion in the first place rested on this connection, it will be useful to see it more clearly in operation.

Before turning to argue about sympathy, though, let me first briefly address how I understand this emotion. In a phrase, I understand it as Daniel Batson (2011) has defined 'it,' namely, as 'an other-oriented emotion elicited by and congruent with the perceived welfare of someone in need' (11). Scare quotes decorate the pronoun because Batson himself labels the emotion he has thus defined as 'empathy' (or 'empathic concern'). As is well known, the distinction between sympathy and empathy is fraught and confusing (and often, confused). But I am not proposing to discuss it here.[17] For present purposes, it suffices to observe the crucial aspect of this distinction, which is that sympathy is *other-oriented* by definition, whereas empathy is not. This gives sympathy and (especially) sympathetic action an evident moral quality that empathy—or fear, for that matter—need not have, and often lacks. Batson, for example, rightly emphasises this feature of his definition in explaining why a motive to help the person in need qualifies as 'altruistic' when it arises consequent on having [sympathy] for them.

[17] I discuss the distinction between sympathy and empathy at some length in *EV*, ch. 6.

§4

Let us shift the emotion in view, then, from fear to sympathy. Of course, to do so, we first have to refresh some central details that have been our stock in trade. Thus, whereas the characteristic object of fear is danger, with sympathy it is 'a being in need or distress.' Likewise, while the response characteristically motivated by fear is 'fight or flight,' as we have been abbreviating it, the characteristic response sympathy motivates is some effort to 'relieve the need' or distress that triggers it. Given its moral significance, it also bears emphasis that what sympathy motivates, more specifically, is relieving this need for the other person's sake (the one who suffers it).[18]

On the input side, therefore, sympathy's epistemic contribution is to register the existence of the needs of particular others (and perhaps further, to assist in discerning their nature), as well as to focus the agent's attention on them. On the output side, sympathy's epistemic contribution is to guide the agent in settling on relief of these needs as the appropriate response. These contributions represent the counterparts, for sympathy, of the two generic consequences of triggering an emotion that we extracted from de Sousa's conjecture about emotion at the outset. On both fronts, I take it, it is evident how the operation of sympathy can be understood as contributing to an agent's *moral* reasoning, rather than as simply contributing to her generic practical reasoning.

---

[18] Evidence that sympathy is genuinely other-oriented can be found, for example, in what happens when the agent's initial effort to relieve the person's need fails for one reason or another. Where the initial effort was motivated by sympathy, such failures typically motivate some further remedial effort on the agent's part. Psychologists rely on this fact to discriminate sympathetic motives empirically from various self-oriented motives for relieving another person's distress. See, e.g., Batson (2011: 117-21).

However, as we have also seen, any such wider contribution a given emotion may be supposed to make depends, among other things, on its being the case that the response it motivates is a morally appropriate response (under the circumstances); and there are various ways in which an emotion—here, sympathy—can fail this necessary condition. For example, it is well known that sympathy is subject to out-group bias of various kinds.[19] Many people may be reasonably alert and responsive to the needs of others when these others are fellow members of some *in-group*, while largely failing to be either alert or responsive to the needs of out-groups.

For concreteness, consider someone whose sympathy is biased in just this standard way, whom we may call Pharisee. One day, Pharisee encounters someone from an out-group who has been beaten by bandits and left for dead at the side of the road. As usual, Pharisee has no sympathy for the man and, ignoring his pleas, walks right on by. No doubt, Pharisee's response is morally inappropriate. Indeed, we may suppose, it is morally wrong. So described, the operation of his sympathy on this occasion (i.e., its non-operation) is plainly morally defective. But what needs greater notice is that this defect infects the operation of Pharisee's sympathy elsewhere, too, even in relation to in-group others. Despite the fact that he is (ex hypothesi) reliable at noticing and responding to the needs of in-group others, the epistemic contributions of Pharisee's sympathy are not *morally* reliable, and hence make no contribution to his moral reasoning. His sympathy is not morally reliable because it does not register or respond to the morally relevant feature of situations in which in-group others are in need—it targets 'in-group need,' rather than 'need' per se.

[19] For an overview of relevant empirical work, see Echols and Correll (2012) and Cikara, Bruneau, and Saxe (2011). (Notwithstanding the variant terminology, this work does engage the emotion of interest to us.)

One possible remedy for this moral imperfection of generic sympathy is to add 'out-group needs' to the eliciting conditions for an individual's sympathy disposition.[20] Evidently, this remedy presupposes that sympathy is suitably plastic in relation to its eliciting conditions, which returns us to our primary question. Without dwelling on the possibility, it should be acknowledged that the addition of out-group needs to sympathy's eliciting conditions may turn out to be only a partial remedy, even for the particular imperfection of out-group bias. For at least in principle, the needs of others in out-groups may reliably trigger someone's sympathy without thereby being registered for the subject as 'calling for relief' or (separately) without the subject's being motivated to relieve these needs. In other words, the input and the output sides of sympathy may come apart to varying degrees in cases of out-group need. On the other hand, insofar as its two sides continue to operate in tandem, adding 'out-group need' as an eliciting condition for someone's sympathy will have the bonus effect of adding 'relief of out-group needs' to the *responses* motivated by their sympathy, too.

My position is that sympathy is not only significantly plastic in relation to its eliciting conditions, but that this plasticity extends to the addition of out-group needs. I shall advance two arguments for this conclusion, beginning with a fixed culture argument. Recall that the fixed culture model of plasticity turns on the availability of some idiosyncratic course of development, by means of which certain default features of the emotion can be changed. Where my previous example of the requisite mechanism was classical fear conditioning, my new example will be Samuel Gaertner and John Dovidio's (2000) 'common in-group identity model.'

[20] In reprising my arguments about plasticity in the case of sympathy, I concentrate on the possibility of *adding* eliciting conditions and neglect the possibility of subtracting them. This is purely for reasons of economy. For a more even-handed discussion, see *EV*, ch. 7. See also note 8.

Now, strictly speaking, Gaertner and Dovidio's mechanism does not target sympathy directly, but rather out-group bias more generally. Moreover, it does not aim to eliminate out-group bias per se. Instead, it aims to harness in-group favouritism to good effect (i.e., to harness pro in-group bias, which differs from anti out-group bias, at least intensionally). Specifically, it aims to *extend* in-group favouritism *to* individuals formerly categorised as members of an out-group, thereby eliminating the subject's bias 'against' *them.*[21] Thus, for our purposes, the question is not merely whether Gaertner and Dovidio's mechanism works to reduce out-group bias, but whether any such reduction in someone's out-group bias actually allows or enables their sympathy to operate with a wider (and so, morally improved) scope.

Together with colleagues, Gaertner and Dovidio (2000, ch. 4) conducted various studies in which some intervention succeeded in emphasising a common group identity among individuals belonging to separate, potentially competing groups and where this induced salience of a common identity was then associated with a clear reduction in intergroup bias. Sometimes their interventions called attention to preexisting superordinate group memberships and other times they introduced new factors that were then perceived to be shared across groups.

Let me describe one of their studies, which took place at a college football stadium before a game. Black and white students approached fans of both the home and away teams prior to their entry into the stadium and asked if they would be willing to be interviewed about their food preferences. Interviewers approached fans of the same sex as themselves and systematically varied

[21] To the extent that some such technique manages to reduce out-group bias extensionally, that is presumably a moral improvement. If a subject eventually thereby achieves the limiting case of being extensionally free from out-group bias, does it remain morally objectionable if his relations with others are still mediated by an in-group category? That is an interesting question, though I shall not pursue it.

whether they wore a home or an away team hat. When black interviewers shared a common university identity with white fans, they obtained reliably higher compliance (59 percent) than when they did not (36 percent).[22] To quote Gaertner and Dovidio's summary, 'these findings offer support for the idea that outgroup members can be treated especially favourably when they are perceived to also share a more inclusive, common ingroup identity' (2000: 63).

Finally, some of Gaertner and Dovidio's studies support a stronger conclusion, which goes beyond the generic reduction of out-group bias and reaches the more specific question about the operation of *sympathy* in relation to former members of out-groups.[23] For example, in one such study, white students viewed a video of a black student being interviewed (Dovidio et alia 2010). The interview was designed both to convey a positive impression of the black student and to elicit an expression from him of either a common group membership (same university), compared to the white subjects, or a different one (race). Later in the interview, the student described a problem he was experiencing, as well as a related task with which he needed help. As part of the experiment, measures were taken of the subjects' attitudes towards the interviewed student, of their [sympathy] for his problem, and of their willingness to help with the task. Across all three measures, responses were 'much more positive' in the common membership scenario than in the different membership scenario (399). Furthermore, the effect of the common group representation

[22] White interviewers obtained similar levels of compliance from white fans, whether they shared a university identity (44 percent) or not (37 percent). White fans also exhibited similar levels of compliance with interviewers wearing a *rival* hat, whether the interviewer was white or black.

[23] Other studies supporting similar conclusions about an improvement in [sympathy] from interventions on out-group bias (albeit using different techniques) are reviewed by Batson and Ahmad (2009). Malhotra and Liyanage (2005) observed effects on [sympathy] a full year following the intervention on out-group bias.

on helping was 'significantly mediated' by [sympathy], whereas there was 'no significant mediating effect' by the subjects' attitudes towards the interviewee (399).

It therefore turns out that some of the evidence anchoring my fixed culture argument about the plasticity of sympathy exceeds our expectations. I originally framed that argument as aiming to exhibit that out-group needs could be added as an eliciting condition for sympathy. But what we have found, as an extra, is that relief of out-group needs (i.e., helping) can also be added to sympathy's *responses* as a concomitant effect of having reduced generic out-group bias. With respect to out-group needs, in other words, sympathy is plastic on both its input and output sides.

§5

To round our discussion out, let me now advance a *variable* culture argument for the conclusion that out-group distress can be added as an eliciting condition for sympathy. Unlike with my previous fixed culture argument, this argument about plasticity will not spill over the line onto the output side of sympathy. We thereby get a neat reversal of the mapping the variable culture argument followed in the case of fear, where it had the task of exhibiting plasticity on the response side.

There is another noteworthy difference between the two versions of the variable culture argument. When it was applied to fear, the salient dimension along which the different cultures varied was explicitly related to their respective *emotional* contents—specifically, to their variant display rules—and the destination culture featured in the argument was deliberately selected on this basis. By contrast, the operative dimension of variation in the present case will be the racial composition of different societies—in other words, not a 'cultural' difference at all. In consequence, it is not entirely clear what drives the outcome (nor shall we try to figure it out).

However, given that the issue simply concerns the plasticity of an emotion, we do not really need to understand the mechanism by which the relevant feature of the emotion has been changed.

In this discussion of the plasticity of sympathy's eliciting conditions, I shall make use of a surrogate marker for the actual triggering of sympathy in this or that individual. Among other things, this allows us to approximate a more precise grip on the triggers, as opposed to relying on self-reports or inferences from characteristic behaviour. Recent neuroimaging studies have explored the neural structures implicated in affective arousal, which attends the processing of negative emotional stimuli. A number of these studies investigate neural activity in response to perceived pain in others. 'Perceiving the pain of others activates brain regions in the observer associated with both somatosensory and affective-motivational aspects of pain' (Contreras-Huerta et alia 2013: 1). These 'empathic neural responses,' as they are called, will be my surrogate for the triggering of sympathy ['empathic concern'].

One of the well-established findings from this research is the existence of racial bias in empathic neural responses, which is an instance of out-group bias as previously discussed. For example, Xu and colleagues (2009) showed video clips to seventeen Chinese and sixteen Caucasian subjects while they were being scanned using fMRI imaging. On the video clips, Chinese and Caucasian models, each with a neutral facial expression, were depicted receiving either a painful stimulation (needle penetration) or a non-painful stimulation (Q-tip touch) in the cheek. Across all subjects, viewing the painful stimulation applied to an in-group face (same race) induced increased activation in the anterior cingulate cortex (ACC) and inferior frontal/insular cortex, as compared to viewing non-painful stimulation to in-group faces. However, when viewing painful stimulation applied to an *out-group* face (different race), the empathic neural response in

the ACC 'decreased remarkably' and this effect 'was comparable in Caucasian and Chinese subjects' (828).[24]

Interestingly, once the researchers from this laboratory had established a racial bias in empathic neural responses, they became curious about whether it was inevitable or was instead susceptible to change. As part of their investigation of its susceptibility to change, they conducted a study that happens to supply more or less custom-made materials for my variable culture argument. Zuo and Han (2013) recruited twenty Chinese adults who had all developed to maturity in a majority-Caucasian culture (some had been born and grown up in a Western country and others had emigrated to one at an early age). Zuo and Han otherwise adhered to the protocol of Xu et alia (2009). What they found was that, unlike in that first study, their subjects exhibited no racial bias in empathy whatsoever:

> [O]ur participants showed significant empathic neural responses to the suffering of both same-race and other-race individuals in the brain region involved in empathy. The interaction analysis did not reveal significant difference in empathic neural responses to Asian and Caucasian models (2013: 42).

Using these materials, we can argue as follows. Let China be the original culture (C1) and Canada be the destination culture (C2).[25] Let E1 designate a set of eliciting conditions for sympathy that is restricted to Chinese others (i.e., in-group members) and E2 a set of eliciting conditions for sympathy that augments E1 by the addition of Caucasian others (i.e., members of one out-group). As Zuo and Han's (2013) study illustrates, we do not have to restrict the transportation modalities between C1 and C2 to cross-cultural adoption, but can expand them to

[24] As Xu and colleagues clarify, 'the ACC mainly contributes to the affective component of empathy' (828).

[25] The three Western countries in Zuo and Han's (2013) study were Canada, the United States, and the United Kingdom.

include emigration of parents with young children,[26] which multiplies the real world instances. Thus, consider an infant in China, Yi. If Yi grows up in China, the eliciting conditions for her adult sympathy disposition will be E1, which means that its being triggered by other human beings in objective distress is subject to a racial bias. Alternatively, if Yi moves to Canada and grows up there instead, the eliciting conditions for her adult sympathy will be E2, which means that *its* being triggered by other human beings in objective distress will be subject to *less* racial bias than it would otherwise have been (i.e., it will not be biased against Caucasians). Insofar as E2 improves on E1, by removing some of E1's moral imperfection, Yi's sympathy has not only been shown to be developmentally plastic, but its plasticity allows for its own (partial) perfection.

[26] Indeed, we could expand them further to include emigration of parents-to-be before they have children, which does not materially affect the point. (Thirteen of Zuo and Han's subjects fit this description). Cao et alia (2015) conducted a similar study, and obtained similar results, with subjects who had all emigrated from China themselves—in their case, to Australia and at older ages.

References

Batson, C. D. (2011) *Altruism in humans*. New York: Oxford University Press.

Batson, C.D. and Ahmad, N. (2009) "Using Empathy to Improve Intergroup Attitudes and Relations," *Social Issues and Policy Review* 3(1): 141-77.

Cao, Y., Contreras-Huerta, L., McFadyen, J., and Cunnington, R. (2015) "Racial Bias in Response to Others' Pain is Reduced with Other-Race Contact," *Cortex* 70: 68-78.

Cikara, M., Bruneau, E., and Saxe, R. (2011) "Us and them: Intergroup failures of empathy," *Current Directions in Psychological Science* 20(3): 149-53.

Contreras-Huerta, L., Baker, K., Reynolds, K., Batalha, L., and Cunnington, R. (2013) "Racial Bias in Response to Others' Pain," *PLOS ONE* 8(12): e84001.

D'Arms, J. and Jacobson, D. (2000) "The Moralistic Fallacy: On the 'Appropriateness' of Emotions," *Philosophy and Phenomenological Research* 61: 65-90.

Dennett, D. (1987) "Cognitive wheels: The frame problem in AI." In Z. Pylyshn (ed.) *The Robot's Dilemma: The Frame Problem and Other Problems of Holism in Artificial Intelligence*. Norwood, NJ: Ablex Publishing.

Dovidio, J., Johnson, J., Gaertner, S., Pearson, A., Saguy, T., and Ashburn-Nardo, L. (2010) "Empathy and Intergroup Relations." In M. Mikulincer and P. Shaver (eds.) *Prosocial Motives, Emotions, and Behavior: The Better Angels of our Nature*. Washington, D.C.: American Psychological Association, pp. 393-408.

De Sousa, R. (1987) *The Rationality of Emotion*. Cambridge, MA: MIT Press.

Echols, S., and Correll, J. (2012) "It's more than skin deep: Empathy and helping behavior across social groups." In J. Decety (ed.) *Empathy: From bench to bedside*. Cambridge, MA: MIT Press.

Ekman, P. (2007) *Emotions Revealed*, 2nd edition. New York: Henry Holt.

Faucher, L. and Tappolet, C. (2008) "Facts and Values in Emotional Plasticity." In L. Charland and P. Zachar (eds.) *Fact and Value in Emotion*. Amsterdam: John Benjamins.

Gaertner, S. and Dovidio, J. (2000) *Reducing Intergroup Bias: The Common Ingroup Identity Model*. New York: Routledge.

Griffiths, P. (1997) *What Emotions Really Are*. Chicago: University of Chicago Press.

LeDoux, J. (1996) *The Emotional Brain*. New York: Simon and Schuster.

Malhotra, D. and Liyanage, S. (2005) "Long Term Effects of Peace Workshops in Protracted Conflicts," *Journal of Conflict Resolution* 49(6): 908-24.

Öhman, A. (2010) "Fear and anxiety: Overlaps and dissociations." In M. Lewis, J. Haviland-Jones, and L. Barrett (eds.) *Handbook of Emotions*, 3rd ed.. New York: Guilford.

Phelps, E. and LeDoux, J. (2005) “Contributions of the amygdala to emotion processing: From animals models to human behavior,” *Neuron* 48: 175-87.

Sreenivasan, G. (2020) *Emotion and virtue*. Princeton: Princeton University Press.

Tappolet, C. (2016) *Emotions, Values, and Agency*. New York: Oxford University Press.

Xu, X., Zuo, X., Wang, X., and Han, S. (2009) “Do You Feel My Pain? Racial Group Membership Modulates Empathic Neural Responses,” *Journal of Neuroscience* 29(26): 825-29.

Zuo, X. and Han, S. (2013) ‘Cultural Experiences Reduce Racial Bias in Neural Responses to Others’ Suffering,” *Cultural Brain* 1(1): 34-46.