# On salience, confabulation, and emotion's reliability

Samuel Lepine

**On salience, confabulation, and emotion's reliability**

Samuel Lepine

**Abstract** De Sousa notoriously insisted on the fact that emotions focus our attention on some information that they make salient. Thanks to this property, they foster the search of reasons, and notably of justifications for our evaluative judgments. But this property has also side-effects since it gives way to various rationalizations and confabulations. When one confabulates, one produces a justification of one's emotion that one genuinely assesses as a good justification, although it is generally not. This raises the problem of emotions' unreliability: how can we trust our emotions if they can lead us in errors of which we can hardly be aware? In this paper, I review various solutions that have been proposed to address this problem. I first examine what I understand to be De Sousa's own solution. I then consider a solution that calls for a form of evaluative understanding, notably proposed by Elgin and Brady. Finally, I suggest another kind of solution, which consists less in distrusting our own emotions, than in a critical assessment of the motivational states from which our emotions derive.

## 1. Introduction

Ronald De Sousa insisted many times on the fact that emotions focus our attention on some information that they make salient. Thanks to this property, they foster the search of reasons, and notably of reasons for our evaluative judgments. But this property has also side-effects since it gives way to various rationalizations and confabulations about our emotional responses. Confabulation seems to be particularly problematic for the justification of our emotions. Indeed, when we confabulate, we produce justifications of our emotion that we genuinely assess as good justifications, although they generally are not. This raises the problem of emotions' unreliability: how can we trust our emotions if they can lead us in errors of which we can hardly be aware? The problem of emotions' unreliability, thus understood, is not that emotions in general are not reliable, but that we can never be sure that an emotional justification is the output of a reliable epistemic process, and not an output of confabulation. As a consequence, there is always a suspicion about the epistemic reliability of the cognitive processes that generate these justifications.

In this paper, I review various solutions that have been proposed to address this problem. I first examine what I understand to be De Sousa's own solution. This solution amounts to say, if I understand De Sousa correctly, that emotions do not foster our epistemic

evaluative knowledge, but only a specific kind of practical knowledge about values. I then consider another solution that calls for a form of evaluative understanding that would be that of a virtuous agent, notably proposed by Elgin and Brady. According to this approach, we can never take our emotions at face value, but we need to exercise a critical reflection on them, so that we can integrate the evaluative judgments they elicit into our evaluative knowledge. Finally, I suggest another kind of solution, which consists less in distrusting our own emotions, than in a critical assessment of the motivational states from which our emotions derive. In doing so, I also propose to reappraise the difficulties that rationalization and confabulation are likely to pose for the justification of our evaluative judgments.

## 2. Emotions and confabulation

Emotions focus our attention on certain information that they make salient. In doing so, they play a determining role in the formation of our beliefs. Being annoyed by a friend's remark, we come to think that he had malicious intentions towards us. As De Sousa notes in *The Rationality of Emotions*: 'emotions are species of determinate patterns of salience among objects of attention, lines of inquiry, and inferential strategies' (1987, 196). In this way, emotions seem to play an intermediary role between desires and beliefs, 'setting the agenda' for these ones: 'they ask the questions that judgment answers with beliefs and evaluate the prospects to which desire may or may not respond' (1987, 196). Thus, emotions are not only a 'source' of reasons, but they also play the 'the role of arbitrators among reasons' (1987, 198-200). They are a source of reasons in that they make us sensitive to certain particular information (like the smile of our friend, the way he looked at us, but also some memories of our past exchanges for instance). And they arbitrate reasons because they will motivate us to favor some reasons rather than others in the formation of our beliefs, by influencing upstream the salience of the information to which we will be sensitive.

These patterns of salience are themselves governed, according to De Sousa, by 'paradigmatic scenarios'. Those scenarios determine the typical situations in which some specific objects are tied to specific emotions (1987, 182), and they are determined by our biography, and by a combination of natural and cultural parameters. Paradigmatic scenarios also set at the same time the 'formal objects' of these emotions, that is, the conditions under which an emotional response may be considered as appropriate or correct regarding its

object[1]. The appropriateness of emotions, according to De Sousa, seems to depend on the fact that they represent accurately who we are (this is the requirement of *authenticity*) and the objective properties of their object (requirement of *truth*), as well as on their being consistent with our other mental states through time (requirement of *rationality*). There are then three dimensions of correction, which are nevertheless 'intertwined' and merged 'into a single complex requirement' (De Sousa 2011, 69-84). It seems appropriate, for instance, to rejoice about the success of my daughter at a tennis game, because her success is an objective fact, and because I have a particular relation with her which makes my rejoicing consistent with my other affective attitudes towards her and with my biographical background. To be correct, then, an emotion should not only correctly appraise its object, but it must do it consistently with some of our motivational states, like our desires, sentiments (love and hate, for instance, which are not episodic states like emotions but long-lasting dispositional states), and preferences.

Thanks to their patterns of salience, emotions help us to resolve a thorny cognitive problem, namely the 'frame problem' (1987, 192), which boils down to determining what kind of information will be relevant for a judgment or a decision, among the indefinitely open set of information at our disposal. An emotion, indeed, 'limits the range of information that the organism will take into account' (1987, 195), so that it allows us to attach more importance to certain information when judging or deciding. More generally, emotions thus favor the development of evaluative knowledge, provided that they are justified and that they generate justified evaluative beliefs about their objects (Tappolet 2000; Deonna & Teroni 2012). If my anger is justified by its object (let's say I am angry at someone who acted disrespectfully, passing me in the supermarket line), then it can also justify the judgement that this object is offensive.

The problem is that by favoring the selection of information in this way, emotions may also favor some cognitive biases, starting with the confirmation bias. As Prinz pointed out, this bias 'helps to explain the attentional effects of emotions and interactions with cognitive processing' (2004, 243), like the fact that sadness draws our attention on flaws, that fear draws our attention on threats, etc. But this bias can also lead us to hold onto our prior beliefs, without assessing them. And not only do we keep these beliefs, but we will try to find 'good' reasons to keep them, even though we should not (Brady 2013, 162). Our emotions, indeed, regularly exert distorting effects on the formation of our beliefs and judgments to 'justify

[1] I will use 'appropriate', 'fitting', and 'correct' interchangeably in this paper.

themselves', as Malebranche once said (Malebranche 1979, 575). They push us to invent all sorts of stories as rationalizations, and they thus generate justifications that bias our judgments. Being angry, we will tend to focus on the insignificant details that could justify our anger.

The same mechanism that allows us to solve the frame problem, then, is also likely to generate a kind of axiological irrationality. This emotional heuristic can both allow a quick distribution of our attentional resources but can also push us to seek justifications for inappropriate emotions. De Sousa points out the same problem in *Emotional truth*: 'When acting "under the influence of emotion" – leaving aside, for the moment, the precise meaning of that expression – we are prone to come up with rationalizations and confabulations' (2011, 35). If the link between emotion and rationalization has been well described so far in philosophy of emotions, the link between emotion and confabulation may have not yet received all the attention that it deserves, especially when it comes to the justification of emotions. I would therefore like to add some considerations about confabulation here, in order to point out the problems it may pose for the justification of emotions, and more generally for their reliability.

Confabulation is commonly defined as an attitude that consists in producing ill-grounded narratives or adopting ill-grounded beliefs, without being aware that these narratives or beliefs are ill-grounded (Bortolotti, 2018). When we confabulate, we generally hold as true, and in good faith, beliefs that are either false or ill-grounded. Hirstein (2009) defines confabulation as a kind of 'epistemic overconfidence', since it leads us to develop beliefs which are often not supported by the relevant evidence. In a nutshell, confabulation is a particularly acute form of rationalization where one cannot even realize that one is rationalizing.

There are different forms of confabulation, some pathological (in anosognosia, or in Korsakoff syndrome, for instance), and some more trivial and ordinary (when one tries to explain one's choices in a supermarket), but in both cases, confabulation implies an effort to restore coherence between our beliefs themselves, or between our beliefs and our emotions, or even between our beliefs and our other motivational states such as our desires, preferences, or attachments. A person may confabulate to explain why she does not want to move her arm (whereas she just cannot and she's unaware of that), or why she desires this brand of pasta (whereas she has just been exposed to an intensive commercial campaign). We are probably prone to confabulate also when we try to interpret our behavior and speak about ourselves without knowing the true causes of our behavior (Nisbett & Wilson 1977).

It is tempting to suspect that there is a special link between affective states and confabulation. Our desires, for instance, lead us to embellish the properties of their objects, whereas frustration leads us to change our beliefs about the desired object, without any change in this one, as in the fox and the grapes fable (Elster 1983). Joy may lead us to neglect the reasons that we have of being worried about an incoming danger, and anger often prompts us to overestimate certain information, in order to find reasons in its favor, even though it was obviously inappropriate to be angry. Disgust also pushes us sometimes to invent reasons to condemn some behaviors, even when there is nothing wrong with these behaviors, as it has been shown in an experiment where Wheatley and Haidt (2005) conditioned students to experience disgust when hearing the word 'often'. Some of these students, indeed, disapproved Dan, a fictitious character, when they were told that he 'often' organizes meetings between students and professors to foster discussions between them. When they were asked why they condemned Dan, one of them said that "it just seems like he's up to something", while another considered him as a 'popularity-seeking snob'[2].

Literature is also full of emotional confabulations where subjects try to justify inappropriate emotions, precisely because they are under the grip of these emotions. In *The Bostonians*, Henry James portrays a character, Olive Chancellor, who is particularly prone to confabulation. Olive is a feminist activist, who places all her hopes in a young orator, Verena Tarrant, to spread feminist ideas. It is quite clear that Olive also loves Verena passionately, so that she is jealous of every person approaching her. But Olive always finds good reasons for her jealousy, and for every negative emotion that comes along with jealousy. In fact, she never really sees herself as experiencing jealousy, nor any negative emotion that may be linked to her jealousy. Rather than admit the fact that she hates Verena's parents, for instance, she prefers to tell herself that she despises them not 'as individuals' but 'as a type, a deplorable one, a class that, with the public at large, discredited the cause of the new truths'. Similarly, she always has good reasons to despise the men with whom Verena befriends, even when – like Matthias Pardon – they share with Verena traits that should move her ('and it is curious that those qualities which he had in common with Verena, and which in her seemed to Olive romantic and touching […] availed in no degree to conciliate Miss Chancellor'). Henry Burrage, another suitor for Verena, finds no grace in Olive's eyes either, although he appears to be a very respectable and charming man. As she realizes this, Olive finds herself forced to seek some reasons to despise him: 'This was painfully obvious when the visit to his rooms

[2] It should be noted that Dan's action was only slightly considered as morally bad, since the mean rating of its moral wrongness was 14 on a scale from 0 to 100 (where it was only 2.7 for the control group).

took place; he was so good-humoured, so amusing, so friendly and considerate (…) that Olive, part of the time, sat dumbly shaking her conscience, like a watch that wouldn't go, to make it tell her some better reason why she shouldn't like him'.

Olive's emotions thus follow a kind of paradigmatic scenario according to which they are supposed to detect the many threats to both her feminist ideals, and to her special relationship with Verena. They put in salience the properties that may frame her deliberations in these terms. It is also likely that she is confabulating since she really tries to be faithful to her ideals. She does not show bad faith in the sense that she doesn't appear to be clearly aware that her emotions and the resulting judgments are ill-grounded. In a sense, she's just striving to build consistency in her overall evaluative perspective, comprising her emotions, judgments, and motivational states. And it is likely that, in that regard, Olive Chancellor is quite an ordinary confabulator. Just like Olive, we all try to find consistency in our evaluative perspective, and we are prone to rely on our emotions to achieve that end.

But it is also clear that Olive seems to have access to the very information that could lead her to revise her judgments, at least concerning Matthias Pardon and Henry Burrage. And this is precisely what distinguishes ordinary confabulations from pathological confabulations. In pathological confabulations, subjects just do not realize that they are making up ill-grounded narratives since they are deprived of certain sources of information, due to brain damage most of the time. So, their confabulations clearly appear as far-fetched. On the contrary, in ordinary confabulations, we have access to the relevant information for our judgments, but we just come to ignore it, because our emotions put in salience other information. Our confabulations, therefore, do not appear as fanciful, but on the contrary as plausible (Bortolotti 2018). We put aside some information, or see that same information under another light, such that we can reconcile it with our evaluative perspective. But even in cases of ordinary confabulation, we remain unaware of the fact that our emotions are motivating us to manipulate information to make it fit our evaluative perspective.

It is then likely that some emotions lead us regularly in error about values, and this would dangerously undermine their reliability. For our emotions to be considered reliable, we would need to be sure that they do not deceive us in this way. However, their link with confabulation leaves open the possibility that they may often give us the impression of being justified without actually being so. The reliability of emotions problem, then, is not that they could be 'volatile', 'instable', or 'fallible', as it is sometimes thought (Elgin 2008, 37). On these different aspects, they do not seem more nor less reliable than perceptions or judgments. The problem is rather that we cannot even be sure that they are a reliable source of axiological

knowledge since they can always implicate a form of confabulation. In the grip of anger, we may be led to think that a remark is offensive, whereas there was really no offense in it. If so, then there is at least a tension between this conclusion, and De Sousa's view – which is also now widespread among philosophers of emotions – that emotions favor the acquisition of reasons, and mostly of reasons for judgments of value.

One possible way to solve this dilemma would be to argue that emotions favor the acquisitions of practical reasons, without giving us any good reasons from an epistemic point of view. Let's call this the 'practical solution' to the problem of emotion's reliability. This seems to be the solution that De Sousa has in mind (2011, 81), following the analysis of Karen Jones who argues that emotions make salient properties that are relevant for our practical deliberations, but which could be also epistemically detrimental. More exactly, some emotions could fail to generate true evaluative beliefs, while nevertheless being rational from a practical point of view, in the sense that they would favor a 'rational framing' (Jones 2004, 341) of a situation, by making us sensitive to considerations that will give us relevant reasons to take our decision. I may, for instance, feel distrustful of a person and then avoid some harm that she might really cause me, even if my distrust is not epistemically justified by the attitude of that person. My distrust would nevertheless be practically rational, according to Jones, if there was only a small chance that this person was indeed toxic, and that because of my biographical background, I would be particularly keen to avoid such relationships.

The point, then, would be that we don't need our emotions to be epistemically correct, we just want them to give rise to appropriate actions regarding the things we care about. So that even if we come to confabulate in justifying an emotion, then, this would not necessarily be a bad thing, if it helps us to choose the right action given one of our motivational states. But this move seems to amount to give up on the idea that our emotions may justify our evaluative judgments. It's not clear, however, that this solution is completely satisfactory. It can be problematic to detach in this way the practical fittingness of an emotion from its epistemic fittingness. We do not want to be afraid of anything and everything just because it may save our life, for instance. We do not want either our emotions to be practically appropriate out of sheer luck. We want to be afraid of things whose properties manifest a real danger. And it is quite a common move to assess the correctness of our emotions, before seeking to know what we should do on the basis of these emotions (Skorupski 2010). I have good practical reasons to ask for excuses, for instance, because I have good epistemic reasons to be angry (the remark was really offensive). Our epistemic reasons, to say it otherwise, ground our practical reasons.

But even if, for the sake of the argument, we agree to detach our practical reasons from our epistemic reasons, it remains true that our emotions could give us the illusion that some considerations are practically relevant when, on the contrary, they would be detrimental. Olive Chancellor always finds good reasons for her negative emotions, but it's not clear that it is good for her, nor for the cause she defends. Confabulation, then, can also lead us astray when it comes to our own interest and motivations. So, the practical solution to the problem of emotion's reliability would not really do the job as such. In the next section, I propose to consider epistemic solutions to this problem.

## 3. Facing emotions' unreliability with understanding

Let's sum up: as they generate confabulation, emotions prevent us from accessing reasons which are nevertheless relevant for our beliefs and our motivational states. Anger prevents me from discovering that I have a reason to think that my anger is inappropriate, indignation pushes me to find reasons to condemn behaviors that have nothing to condemn, and so on. Our emotions selectively focus our attention to what is typically likely to confirm them and exclude other available contextual elements. They can thus generate complacent, even false reasoning, and blind us to the reasons why we experience them, as well as to the reasons we would have for not experiencing them. In some cases, indeed, our emotions can lead us to invent reasons to rationalize them that we see as relevant justificatory reasons, whereas they are not. It is even highly plausible, in this respect, that we are sometimes prisoners of paradigmatic scenarios that calibrate our emotions: having grown up in a patriarchal and honor-based culture, I see all the reasons I would have to be jealous about my partner's behavior, although none of these reasons make my jealousy justified. Can we seriously maintain, then, that emotions can justify our evaluative judgments? Are they reliable enough to play this epistemic role?

Against this line of reasoning, it could be argued that, in the cases mentioned above, the emotions are themselves unjustified. Olive's disgust is never justified: it is the product of bad a priori reasons that prevent Olive from properly assessing people's worth. And since this emotion is unjustified, it will generate unjustified judgments. But if Olive were more lucid about her own epistemic situation (that is of her own emotional biases), she would stop being blind in this way. The problem, however, to paraphrase Goldie (2008, 159), is that 'misleading emotions' in this kind of situation tend to mask their 'own misleadingness', so that it is unlikely that we could become aware of this deception.

Although Goldie defends this view about another type of emotion (those that he believes are the result of an 'evolutionary mismatch' such as xenophobia or jealousy), what he says about 'misleading emotions' seems particularly relevant for emotions that generate confabulation. Indeed, he argues that deceptive emotions tend to 'distort the epistemic landscape' (Goldie 2008, 159). The angry person, as we saw, will tend to see certain details as clear evidence of the offense he has suffered. The emotion then 'masks its own misleadingness', according to Goldie, because calm and deliberative reflection will hardly alter the emotional assessment. The emotion, indeed, has already undermined deliberation unbeknownst to us, by discarding information and evidence relevant to its own correction and, conversely, by giving too much weight to small details that confirm our initial appraisal.

If emotions can mislead us in this way without our being able to realize it, then this fact is likely to cast doubt on the reliability of all our emotions. For it could be that even when we have the impression that our emotions are reliable, this impression itself is the result of the emotions' misleadingness. If so, we should take our own emotional deliverances with great caution. Following this line of reasoning, some authors argue that emotions do not really give us reasons but constitute a 'material' that will facilitate the acquisition of reasons. According to Elgin (2008, 33), for instance, emotions are thus analogous to iron ore: their value comes from the refinement and cognitive processing that we impose on them, to deepen our 'understanding' of ourselves and of the world. They are epistemic resources whose 'yield' depends on the cognitive capacities that we deploy to exploit them. In a similar vein, Brady (2013, 118) argues that an emotion is a 'proxy', that is a mental state that does not give us genuine reasons, but only '*pro tempore* reasons', which are substitutes and provisory reasons. These reasons are temporary since they are supposed to give way to a more general understanding of our own evaluative perspective.

Elgin and Brady both seem to see understanding as a kind of holistic perspective, where different information, beliefs, and commitments are maintained together in a systematic and coherent way, and where each element is supposed to partly justify other elements. Understanding is then irreducible to knowledge, since it is more than a justified belief about a particular topic (Brady 2013, 137). It is more like an organic whole, which is supposed to be the goal of our emotional appraisals: we do not want only to have justified evaluative judgments, according to this view, but we want to have a web of justified evaluative judgments, emotions, and commitments.

The value of understanding, in this regard, would be superior to the value of evaluative knowledge since it would be more encompassing and richer than the latter. Brady argues that

understanding is the true goal of our emotions, since it incorporates a causal explanation of *why* some properties are tied to some values, and why we should react to the presence of these values in some specific ways (2013, 141). Understanding helps us to answer 'why questions' concerning the justification of emotions, because it relies on a deeper 'insight' of 'genuine' causal relations between facts, values, and actions. It is thus supposed to foster our moral development and social coordination with others, since it is so tightly connected with our ability to act appropriately, on the basis of our ability to recognize values, and what values requires us to do.

According to this approach, then, we should never take our emotional evaluations at face value since they are always likely to mislead us. Every emotion, on this approach, requires a kind of cognitive vigilance, to check if its appraisal fits with our other mental states 'in reflective equilibrium' (Elgin 2008, 48). This kind of approach is notably opposed to perceptual approaches such as the one favored by De Sousa, in which our emotions are supposed to provide us with reliable value perceptions, as long as they are justified. But it is also opposed, more generally, to any view which considers that an emotion can enrich our evaluative knowledge if it is justified. Brady, for instance, argues that we should exercise a critical control over each emotion to achieve an understanding of our own evaluative perspective. The problem, though, according to Brady (2013, 135), is not so much that emotions are unreliable, but that they cannot justify our evaluative beliefs by themselves. This would presuppose a kind of externalism about the justification of emotions, where the justification of an emotion would depend on the properties of its object and would suffice to provide us with genuine evaluative knowledge. Brady's argument is precisely that this is unlikely: we want to understand why the properties of an object – say a bull, to take Brady's example – make it a dangerous thing that we should fear. And we look for an evaluative understanding of our situation, Brady argues, because we are not content to know *that* the bull is dangerous, we also want to understand *why* it is dangerous, and to do this, we need an awareness of the connections between some properties and some values.

It is then the justification process based solely on external evidence that is more generally unreliable, according to this view, because it cannot make sense of the idea that we need to grasp the reasons which make an object or an attitude dangerous, shameful, rejoicing, etc. What understanding opposes is precisely this externalist model of justification. Evaluative understanding, on the contrary, relies more on an internalist view of justification, where one can explain why some properties are supposed to bear certain values (why the bull's behavior is to be understood as a threat), given our grasp of the many relations between natural

properties and values, and the coherence of these relations. This kind of understanding would require an agent with 'virtuous habits of thought and attention' (Brady 2013, 170), who would thus be able to reflect on the validity of his initial emotional appraisal and its integration with his background motivations. If Olive Chancellor were virtuous in this sense, she would not only focus her attention on the properties which are likely to justify her jealousy, but she would also remain open to other considerations, such as the fact that other properties do not justify this emotion, but on the contrary are opposed to it. She would also consider the fact that her jealousy is weighing dangerously on her health and her well-being. As a virtuous agent, she would like to further her evaluative understanding, and so she would reappraise the justifications of her jealousy and her hatred.

An important point is that the virtuous agent is able to exercise a control over her emotional attention thanks to her evaluative understanding. Since she already has an understanding of her situation, and then of the value of some object, she will not be inclined to constantly check her emotional appraisal, and her attention will not be consumed in this way (she knows that flying is safe, so even if she is afraid, she will not trust her fear). She will not be prone to rationalization either, since rationalization is, according to Brady, the result of a lack of awareness of what properties 'constitute genuine reasons that bear on the accuracy of one's emotional construals' (2013, 178). To say it otherwise, the virtuous agent, thanks to her understanding, is not in need to invent fallacious reasons. Evaluative understanding also goes along with a grasp of the conditions under which our emotional appraisal is likely to be unreliable (like when we are in the middle of a breakup, and excessively prone to anger or sadness), and with a 'virtuous regulation of attention' (2013, 185), so that the virtuous agent will remain open to other features that could help her reappraise and discount her first emotional response.

According to Brady, this conception of the virtuous agent is not overly demanding, insofar as it relies primarily on good epistemic habits. Such an agent would almost never trust her emotions, except when circumstances prevent her from exercising critical control over them (because, for instance, she lacks time to do it). The only reasons she would have to rely on her emotions would thus be circumstantial and pragmatic. But in the absence of such circumstances, the virtuous agent should generally not rely on her emotions: 'the more virtuous someone is, the less reliant he will be on his emotional responses to themselves disclose evaluative information' (2013, 188).

Emotions are therefore destined, for the virtuous agent, to play a relatively secondary role since they are never epistemically sufficient to increase our evaluative understanding. Either

the virtuous agent already has an evaluative understanding of his situation, in which case he simply does not need to rely on his emotions; or he does not have a sufficient understanding of his situation, in which case he will subject his emotions to critical reflection. Thus, his emotions can never justify his axiological judgments. Brady concedes, however, that emotions do play such a role in non-virtuous agents: in ordinary life, our emotions serve to justify our value judgments, because we can do little better than rely on them. But ideally, we should subject them to critical reflection to improve their epistemic yield and refine our evaluative understanding.

Is it true then, that this conception of the virtuous agent is not too demanding? On the contrary, I shall argue that it requires abilities that are incompatible with the patterns of emotional salience. The virtuous agent, indeed, must be capable of detaching his attention from the object of his emotion, so as to avoid confabulation. He must demonstrate a certain attentional openness, which will allow him to consider other characteristics relevant to the critical assessment of his emotion, as we saw with the example of Olive Chancellor above. Olive would have to remain open to the properties of certain men, such as Henry Burrage, to revise her general evaluative perspective, and notably her hatred of men. Brady (2013, 185) argues that this kind of openness is not precluded by the narrowing of our attentional focus that some emotions can cause. And of course, when one is experiencing jealousy, one does not necessarily take one's emotion at face value. But this is a trivial fact about our emotional appraisals, and Brady's virtuous agent does more than that: he is able to counterbalance his appraisal with other considerations, precisely because he can in some sense control his field of attention.

In many ways, this view seems incompatible with the nature of emotions. As Goldie (2004, 99) suggests, emotions reinforce our ordinary cognitive biases by making them more resilient: they generate 'idées fixes' that our judgment must embrace. And it is implausible that these biases can be overcome by being virtuous or well-educated. Goldie argues indeed that it is difficult to be aware of these biases since they generally operate covertly. It is also difficult to divert our attention from the very facts that our emotion makes salient, or to make other facts more salient instead of them, precisely because our emotion causes us to ignore these other facts. Another common but illusory solution, Goldie reminds us, is embedded in all these maxims of popular wisdom: 'stop and think; count to ten; bite your tongue; take a deep breath; sleep on it' (Goldie 2008, 162). But again, this assumes that we may be aware of our biases, and this is precisely what is unlikely, especially when one confabulates, being sure to point out in good faith relevant reasons to justify his emotion. Of course, we are often in a

better epistemic situation when we step back from our emotions, especially when we are no longer in their most intense phases, and we can sometimes free ourselves from their distorting effects. But the problem with confabulation, as we have seen, is that it can lead us to think in good faith that we are reasoning objectively when our reasoning is in fact already motivated, and that we are prone to give more weight to some facts or details than to others. So that even stepping back from our emotions may never be a sufficient epistemic guarantee against their distorting effects.

From this point of view, the idea that we could face emotions' unreliability with epistemic virtues seems unlikely, since it is probably impossible to control our attention during the very development of an emotion, or to prevent the appearance of attentional biases. What skill could the virtuous agent exercise that would really limit the power of emotions to direct our attentional focus with patterns of salience? At best, he will just be able to reappraise his emotion, and try to see if it is justified. But I guess that this is what we all do with more or less success, and this is certainly not enough to prevent emotions from directing our attention. The virtuous agent is also supposed to neutralize the rationalization process that comes with certain emotions. And this also seems quite dubious, not only because this kind of process, as we just saw, operates covertly, but also because it is in the nature of rationalization – especially with confabulation – to give us the impression that our justifications are well founded.

So we shall probably not try to solve the problem of emotion's unreliability by an appeal to evaluative understanding nor a critical reflection on our emotion. Not that I want to dismiss completely critical reflection. Once again, I take it that we all, in some ways, try to reappraise our emotions when we can, and we certainly do it with more success when we have good epistemic habits. But what is at stake with emotions' unreliability, once again, is not that all emotions are unreliable. It is that some emotions generate confabulations, and that we cannot ever be sure that we are not confabulating when we try to justify our appraisals. And it just seems implausible that a virtuous agent would be able to detect these epistemic failures when they occur, or to neutralize them when they appear. The solution I would like to turn to in the next section, then, is to focus not on a kind of critical control over our emotions, but on controlling our emotional dispositions. While this may sound as a form of virtue ethics, I will argue that this is not the case.

## 4. Unreliable motivations and emotional ambivalence

In this final section, I would like to propose a solution to the problem of emotion's unreliability which consists not in focusing on the emotions, but rather on the motivational states from which our emotions derive, like our desires, sentiments, or preferences. I suggest that the problem of emotion's unreliability is mainly related to our motivational states. What is wrong with Olive Chancellor is above all her motivations (her hatred of men, her exclusive and possessive love for Verena), which make her blind to certain reasons. She is condemned to look for bad reasons, reasons that will tend to confirm her hatred and her jealousy. This approach thus amounts to postulating that it makes sense to think that there are motivational states that are correct and incorrect, just as there are correct and incorrect emotions. It seems unfitting, for instance, to entertain a sentiment of hate for a person who is nice in every way, or to desire things that have no desirable properties. If so, then we can also think that there are good or bad justifications to assess the fittingness of these motivational states.

The problem, according to this analysis, is not that emotions generate confabulation. In fact, confabulation is a kind of collateral effect of the evaluative function of our emotions and their patterns of salience. Emotions tend to favor the search for relevant reasons for our motivations. When our motivational states are justified, this isn't necessarily problematic. In this case, our emotions will also lead us to look for relevant reasons to maintain and justify these motivations. Let's say, for example, that I love a person who is indeed loveable. My love for this person will cause me to experience all sorts of emotions for her (sadness, joy, anger, etc.), which in turn will cause me to look for reasons in their favor. When I confabulate, in a way, I am simply looking for a coherent story that will allow me to rationalize my emotion, based on certain salient details. Confabulation, in this sense, is a relatively normal rationalization process, resulting from the emotional 'framing' of the situations we encounter, and which can lead us to find good reasons as well as bad reasons in favor of our motivations. The framing of an emotional appraisal focuses our attention on certain elements of a situation that it makes salient, and thus gives them priority on other elements in the formation of our appraisals and the judgments that result from them.

This framing process is entirely independent of our will and our control as it happens, and yet we will form our judgments out of it. Rationalization then intervenes to introduce coherence into the judgments that result from this process. As Cushman (2020, 3) argues, it is then a form of 'inverse planning' whereby we will try to make up beliefs that could account for our emotion. To say it otherwise, rationalization is a process of searching for relevant reasons, by which we extract information from the mechanisms that implicitly guide our choices and behaviors (here, our emotions), in order to make this information useful for

further reasonings. More specifically, we rationalize to make up a story that would have put our emotions or actions in a rational light. 'Rationalization is rational', then, according to Cushman, both in a practical and in an epistemical way. Rationalization is practically rational because it allows the fictional stories that we make up to influence our beliefs, and our subsequent behaviors and reasonings, so that it has an instrumental function: our future decisions will then be made on the basis of the information that we extract from our current and past behavior. But rationalization allows this precisely because it is also epistemically rational, at least in the minimal sense that it allows us to form beliefs about our own behavior which are consistent with our other beliefs about ourselves.

From this point of view, confabulation is just one way to rationalize our emotional appraisals. More specifically, it's a kind of byproduct of the rationalization process that comes with the emotional framing, and that probably happens mostly when an emotion which seems justified from our evaluative perspective, does not find any proper justification in the properties of the emotion's object. As we cannot find any legitimate justification, we are led by the framing process to make up one that could fit with our evaluative perspective. So that, when we are confabulating, we do not really have an epistemic alternative. This also lifts the suspicion, raised earlier, that we could be always confabulating when trying to justify our emotions. We probably confabulate, most of the time, when there is a discrepancy between an emotional appraisal that we think is well founded by our motivations, but that we cannot justify satisfactorily by referring to the properties of its object. From this point of view, emotional confabulation is often 'epistemically innocent' (Sullivan-Bisset 2015), in the sense that it plays at the same time a faulty but beneficial epistemic role. Confabulation, indeed, fills 'explanatory gaps' between some of our mental states, and it generates justifications that are now open to criticism. Since we cannot neutralize the rationalization process, as we saw, the best we can do is to exercise our critical reflection on our confabulatory justifications.

In any case, since we have no real control over the course of the framing process, it seems that we should probably not concentrate our efforts over this one, like an idealized virtuous agent would do. Indeed, the lesson to be drawn from this analysis, I suggest, is that we should remain vigilant about the motivational states from which our emotions themselves derive. Most of the time, our emotions just do what they are supposed to do: they track evaluative properties which are related to our motivational states. We are afraid of not being able to satisfy a desire, we are disappointed that a preference is not fulfilled, and we are indignant at our enemy's behavior. What is at stake, then, is rather the motivations of the individuals who confabulate. Jealous people, like Olive Chancellor, seem to fail to see that their jealousy is

preventing them from flourishing, and that it interferes with the pursuit of other motivations, such as their desire to be happy. What they seem to lack, then, is more a kind of openness to the reasons that they would have to revise their motivations.

In this sense, we should exert a kind of critical reflection on our motivational states, which is far easier than with our emotions, since we have many opportunities to do it calmly, without being in the grip of salience effects. The question then, is to determine when a motivation can be considered as justified. Arguably, there are many ways to assess the justifications of a motivation. For the sake of brevity, I will simply indicate at least three ways to do so. First, as with emotions, it seems that we can assess the justifications of a motivational state with epistemic reasons, depending on its formal object. We can thus ask ourselves if the object of our desire is really desirable, if the object of our love is really loveable, and so on. Is Eric really worthy of hatred? This will depend on various clues such as his behavior, his words, and other objective evidence. Second, we can also ask whether our beliefs about the objects of our motivations are themselves justified. Racist and xenophobic motivations, for instance, are always based on false beliefs. Finally, we can also assess some of our motivations with moral considerations (when they are related to what is morally good or bad, such as a positive moral sentiment towards vegetarianism), or with prudential considerations (when they are related to our well-being, such as my desire to eat chocolate all day).

Based on these various considerations, we can thus ask whether a given motivation is justified and worthy to be endorsed. I readily admit, however, that by following these criteria, most of our motivations will probably turn out to be justified most of the time. It is indeed quite rare, fortunately, that we love people who are not lovable, or that we desire things which are not desirable. The main problems here surely stem more from moral or political motivations, whose justifications are more difficult to assess, notably because they are more subject to controversy.

Having justified motivational states, in any case, is not like developing any particular virtue. It just amounts to check if our motivational states are based on sufficient justifications, and if these justifications are not defeated by certain considerations. It does not require from us that we develop specific skills of thought and attention, or that we become experts in values. But is this sufficient to deliver us from the problems that confabulation creates for the justification of emotions? One could object that we can still confabulate when we have justified motivations, for example by looking for justifications for our partiality towards a

person we love. In that kind of cases, we will still elaborate bad justifications for our emotions, while being convinced that these are good justifications.

But my point is not that confabulation is problematic only when it is linked to inappropriate motivational states. Confabulation, of course, is a problem for emotional justification in general, whether it is linked to appropriate or inappropriate motivations. My point is more that it is still a minor problem, and not one that should be taken as a good reason to distrust our emotional appraisals, since it is in fact the outcome and byproduct of a more general rationalization process that we engage in because of the salience patterns that go along with our emotions. And since our emotions depend on our motivations, I argue that what is at stake in confabulation is more these motivations that we wholeheartedly seek to preserve than our emotional appraisals, which are only doing their job. So the problem with inappropriate motivations is that they tend to lock us in a vicious epistemic circle, where, in the grip of our emotions, we will try to elaborate justifications that will help us to keep them, when, on the contrary, we would need to become aware of their bad justifications.

In this regard, it is worth noting that our emotions can help us to critically examine our motivations, at least when we experience some kind of emotional conflict. This is a point that De Sousa also emphasizes: 'Emotions that are *incompatible*, therefore, are likely to be *felt* as normatively *inconsistent*, felt, that is, as a problem requiring some sort of resolution' (2011, 80). This kind of incompatibility is mostly experienced with cases of emotional ambivalence. As Price (2015, 148-154) points out, indeed, ambivalent emotions seem to indicate conflicts between our motivations, and doing so, they allow us to remain open to the reasons that militate against our motivations. Ambivalent emotions are opposite and simultaneous emotions towards the same object, or the same aspect of a situation. We can thus be annoyed and amused by a noisy child, grateful and envious of a person's generosity, and so on. Of course, it is entirely possible that two ambivalent emotions are simultaneously correct (the noisy child can indeed be both amusing and annoying), but in any case, ambivalent emotions help us to fight against the restriction of the attentional focus that each emotion produces on its own.

Moreover, emotional ambivalence may indicate that we have conflicting motivations, so that we should abandon or revise any of these motivations. I may be torn, for instance, between my desire to live a selfish life based only on the solitary pursuit of pleasures, and the sentiment of love that I have for my newborn child, whose birth was not planned at all. I may thus be inclined to feel despair and joy about this situation, and this felt contradiction will probably lead me to revise one or the other of my motivational states, since I cannot satisfy

both simultaneously. This revision can come in at least two ways: we can realize that one of our motivations was ill-grounded (Olive Chancellor may realize that this is the case with her hatred, when she comes to reflect on her own happiness), or because certain emotions involve 'transformative experiences' (Paul, 2014), in the sense that they modify our motivational set itself: I may prefer not to have a child at $t_1$, and prefer to have one at $t_2$, because the emotional experience of parenthood has modified my preferences in the meantime. This will lead me also to revise my desire to live a selfish life.

Thus, ambivalent emotions can lead us to critically reflect on our motivations and examine which of them are justified or not. But it is also true that it is difficult to revise motivational states that are most central. Central motivational states, indeed, generally enjoy stronger cognitive and affective integration, in the sense that they are connected to a large number of other mental states. Revising them would then also involve revising a number of these other states. Revising my desire to enjoy a selfish life of pleasure is linked to other beliefs related to the value of these kinds of pleasure, to the various forms of alienation that threaten this way of life (including parenthood!). We can then hypothesize that this is precisely an important confabulation factor: the more deeply a desire or a sentiment is entrenched in us, the more cognitive resources we will deploy to maintain it. If this hypothesis is true, then it is with respect to our most visceral motivational states that we should be most vigilant, insofar as they are likely to blind us to the relevance of certain evaluative properties for our other motivational states.

In any case, we should not see rationalization and confabulation as threats to emotion's reliability. On the contrary, they may even help us to criticize our emotional appraisals. It is mainly with regard to our motivational states that we should remain vigilant. But even here, it is worth noting that ambivalent emotions also give us the means to exercise a form of critical reflection, and that they may help us to determine what motivation is worth pursuing.

**References**


Bortolotti, L. (2018). 'Stranger than fiction: costs and benefits of everyday confabulation'. *Review of Philosophy and Psychology*, 9 (2): 227-249.

Brady, M. (2013). *Emotional insight: the epistemic role of emotional experience*. New York: Oxford University Press.

Cushman, F. (2020). 'Rationalization is rational'. *Behavioral and Brain Sciences*, 43, E28.

De Sousa, R. (1987). *The rationality of emotion*. Cambridge: MIT Press.

De Sousa, R. (2011). *Emotional truth*. Oxford: Oxford University Press.

Elgin, C. (2008). 'Emotion and understanding'. In G. Brun, U. Doğuoğlu, et D. Kuenzle (eds.), *Epistemology and emotions* (pp. 33-50). London: Ashgate.

Elster, J. (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Goldie, P. (2004). 'Emotion, Feeling, and Knowledge of the World'. In R. C. Solomon (ed.), *Thinking About Feeling: Contemporary Philosophers on Emotions* (pp. 91-106). Oxford: Oxford University Press.

Goldie, P. (2008). « Misleading Emotions ». In G. Brun, U. Doğuoğlu, et D. Kuenzle (eds.), *Epistemology and emotions* (pp. 149-165). London: Ashgate.

Hirstein, W. (ed.) (2009). *Confabulation: Views from Neuroscience, Psychology, Psychiatry, and Philosophy*. Oxford: Oxford University Press.

James, H. (2000). *The Bostonians*. London: Penguin.

Jones, K. (2004). 'Emotional Rationality as Practical Rationality'. In C. Calhoun (ed.), *Setting the Moral Compass: Essays by Women Philosophers* (pp. 333-352). New York: Oxford University Press.

Malebranche, N. (1979). *De la Recherche de la vérité*. In *Malebranche : Oeuvres, tome I*. Paris: Gallimard.

Nisbett, R. E., & Wilson, T. D. (1977). 'Telling more than we can know: Verbal reports on mental processes'. *Psychological Review*, 84 (3), 231-259.

Paul, L. A. (2014). *Transformative experience*. Oxford: Oxford University Press.

Price, C. (2015). *Emotion*. Cambridge - Malden: Polity Press.

Prinz, J. (2004). *Gut reactions. A perceptual theory of emotion*. Oxford: Oxford University Press.

Skorupski, J. (2010). 'Sentimentalism: Its Scope and Limits'. *Ethical Theory and Moral Practice*, 13 (2), 125-136.

Sullivan-Bissett, E. (2015). 'Implicit Bias, Confabulation, and Epistemic Innocence'. *Consciousness & Cognition*, 33, 548-560.

Tappolet, C. (2000). *Emotions et valeurs*. Paris: Presses Universitaires de France.

Wheatley, T., & Haidt, J. (2005). 'Hypnotic disgust makes moral judgments more severe'. *Psychological science*, 16 (10), 780–784.