Review of Philosophy and Psychology (2021) 12:329–351
https://doi.org/10.1007/s13164-020-00492-8

Emotions and the Problem of Variability

Juan R. Loaiza[1,2]

Published online: 10 June 2020

Abstract

In the last decades there has been a great controversy about the scientific status of emotion categories. This controversy stems from the idea that emotions are heterogeneous phenomena, which precludes classifying them under a common kind. In this article, I analyze this claim—which I call the Variability Thesis—and argue that as it stands, it is problematically underdefined. To show this, I examine a recent formulation of the thesis as offered by Scarantino (2015). On one hand, I raise some issues regarding the logical structure of the claim. On the other hand, and most importantly, I show that the Variability Thesis requires a consensus about what counts as a relevant pattern of response in different domains, a consensus that is lacking in the current literature. This makes it difficult to assess what counts as evidence for or against this thesis. As a result, arguments based on the Variability Thesis are unwarranted. This raises serious concerns about some current empirical theories of emotions, but also sheds light on the issue of the scientific status of emotion categories.

Since Griffiths's (1997) *What Emotions Really Are*, there has been a wide discussion about the natural kind status of emotions. Broadly construed, the question is whether emotions (both «emotion» as a general category, as well as particular emotion categories such as «happiness», «sadness», «fear», and the like) refer to phenomena that can be distinguished independently of our own conceptual framework, presumably in terms of affect programs or at least some discrete pattern at the neural or physiological levels. In more recent years, the focus of this discussion has shifted towards newer empirical evidence challenging a positive answer to this question. Reviews such as Barrett (2006) and meta-analyses such as Lindquist et al. (2012) have attempted to show that emotions are variable phenomena at the neural and physiological levels. This is taken to imply that emotions do not form natural kinds, and therefore that we should reexamine the ways in which we think about emotions scientifically.

✉ Juan R. Loaiza
loaiza.juan@hu–berlin.de

[1] Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany

[2] Escuela de Ciencias Humanas, Universidad del Rosario, Bogotá, Colombia

The claim that emotions are variable phenomena has appeared in the literature with many names. Scarantino (2015) calls it the *Problem of Variability*. Prinz (2004) calls it the *Disunity Thesis*. Under any of these names, this claim is one of the main issues of debate in emotion research. Based on this claim (hereafter the Variability Thesis, or VT), researchers have argued for rejecting traditional theories of emotion such as basic emotion theory or discrete appraisal theories (Barrett 2006), defended skepticism about important past empirical findings (LeDoux 2012), and used it as a platform to construct new theories of emotion (Barrett 2017; Scarantino 2015).

In this paper, I will claim that even though VT has raised important questions and issues for emotion theories, this thesis remains underdefined. As a result, drawing conclusions and basing new theories on VT is problematic, since it is unclear on which empirical findings and under which criteria is the thesis established. This raises serious concerns about some current empirical theories of emotions, but also sheds light on the issue of the natural kind status of emotions and their role in scientific inquiry.

In the first section, I examine the role VT has played in recent discussions about emotions. Following Scarantino's (2015; Scarantino and Griffiths 2011) analysis of VT, I focus on two important consequences drawn from VT in this version. The first is the claim that "variability is the norm", a claim that underlies Barrett's (2017) Theory of Constructed Emotion. The second is a skeptical thesis about the usefulness of folk emotion categories in science, one that has been defended by Scarantino (2012b) and LeDoux (2012). After establishing the importance of VT, in the next section I show why the thesis, as it has been defined in the literature, leads to ambiguity. I show that we can divide VT into several sub-theses, each leading to different theoretical commitments and empirical predictions. I further argue that this ambiguity leads to a lack of consensus regarding what kind of evidence would decide for or against VT, creating an obstacle for comparing different theories of emotions and their empirical predictions.

1 The Variability Thesis

The Variability Thesis (VT), as presented above, can be presented under the following working definition:

Variability Thesis (VT): Emotions are naturally disjoined phenomena.

As Scarantino (2015) and others (Barrett 2006; Prinz 2004) formulate it, it is a thesis directed at basic emotion theory (BET), given that BET (in at least some of its incarnations) expects that emotions—or at least the basic emotions[1]—correspond to discrete neural or physiological patterns. However, this problem is not exclusive to basic emotion theories (Barrett 2006), but rather affects any theory of emotion that expects one-to-one mappings between emotions and some presumed natural.

[1] Ekman thinks that all emotions are basic, thus expecting all emotions to correspond to some pattern, in this case a physiological one. Others such as Izard or Panksepp accept distinctions between basic and non-basic emotions, but identify basic emotions with processes generated by evolutionarily ancient brain systems. (See Ekman 1992; Ekman and Cordaro 2011; Izard 2007, 2009; Panksepp 1998, 2008, 2011; Panksepp and Watt 2011)

Scarantino presents VT (the "Problem of Variability" in his terminology) as the conjunction of two theses:

> *No One-to-One Correspondence (NOC) Thesis*. There is no one-to-one correspondence between anger, fear, happiness, sadness, and so forth, and any neurobiological, physiological, expressive, behavioral, or phenomenological responses.
>
> *Low Coordination (LC) Thesis*. There is low coordination between neurobiological, physiological, expressive, behavioral, or phenomenological responses among instances of anger, fear, happiness, sadness, and so forth. (Scarantino 2015, p. 343)

NOC claims that emotions do not map one-to-one onto processes in the brain or the body, nor to expressions, behavior, or phenomenology. There are two ways in which this can occur. One is that for one given emotion category, there are several associated neural, physiological, expressive, behavioral, or phenomenological responses. The second is that one set of responses is associated with two emotions. I will consider these options in detail when I propose some refinements to this claim below.

The LC thesis, on the other hand, holds that there are no robust correlations between different presumed components of an emotion. In an earlier presentation of this claim, Scarantino and Griffiths (2011) explain:

> Evidence for LC consists of examples of anger, happiness, sadness, surprise, etcetera, that are instantiated in the absence of a coordinated package of physiological, neurobiological, expressive, behavioral, cognitive, and experiential responses. (Scarantino and Griffiths 2011, p. 448).

In this formulation, evidence for LC consists of cases in which one given emotion occurs but there is no set of correlated properties occurring. Barrett (2006), presenting some evidence for this claim, writes:

> Although no single study of emotion has simultaneously measured facial movements, vocal signals, changes in peripheral physiology, voluntary action, and subjective experience, many studies have measured at least two or three of these responses (usually some combination of subjective experience, behavior, and autonomic activity). These studies have reported a range of associations, from modest correlations to no relationship to negative correlations among experiential, behavioral, and physiological measures of emotion. (Barrett 2006, p. 33)

Following this interpretation, LC is a thesis about the correlation between different measurements. LC would be thus established if for a given emotion category, we fail to find that, for instance, skin conductance responses (physiological measure) for anger do not correlate with anger expressions, or that neural activity for sadness fails to correlate with retreat action tendencies characteristic of the emotion (behavioral measure).

Two main consequences have been drawn from accepting VT. First, findings showing lack of correspondence and coordination have raised questions about the natural kind status of emotions. Barrett (2006) explicitly argued that empirical evidence

for variability suggests that emotions are not natural kinds. This move has further motivated skepticism about the use of emotion categories in emotion research. For example, based on VT, LeDoux (2012) has proposed a reconceptualization of how the brain is involved in affective responses, one that makes no use of traditional emotion categories:

> I concur with [Barrett's 2006] conclusion that the foundation of support for the idea that basic emotions, as conventionally conceived, have dedicated neural circuits is weak. This does not mean that the mammalian brain lacks innate circuits that mediate fundamental phenomena relevant to emotion. It simply means that emotions, as defined in the context of human basic emotions theory, may not be the best way to conceive of the relevant innate circuits. (LeDoux 2012, p. 655)

Another important conclusion drawn from establishing VT is that emotions, rather than having neural essences or physiological fingerprints, are constructed. This is the claim at the base of Barrett's *Theory of Constructed Emotion*. In one of the latest presentations of her view, Barrett (2017) explains:

> A constructionist approach to emotion has a couple of core ideas. One idea is that an emotion category such as anger or disgust does not have a fingerprint. One instance of anger need not look or feel like another, nor will it be caused by the same neurons. Variation is the norm. (Barrett 2017, pp. 32-33)

As I hope it is clear by now, VT is at the core of a number of important debates and theories that dominate current emotion research. Moreover, it is taken as an established claim, with an increasing amount of evidence interpreted through its lens. Yet, as I have suggested, the claim is problematically underdefined.

2 Problems with VT

Recall Scarantino's construal of VT as composed of two theses: NOC and LC. So construed, NOC is a thesis about the mapping between emotion categories and patterns of responses, while LC is a thesis about the correlation between these responses. Let us analyze each of these claims in turn.

First, note that NOC admits a subdivision in terms in terms of the type of responses that emotions could correspond to, i.e. neural, physiological, expressive, behavioral, or phenomenological packages. Each of these NOC theses would claim that a given emotion does not correspond one-to-one with a given type of response (e.g. one emotion corresponding to two types of neural response). Thus, we can divide NOC into

> NOC_{Neural}: There is no one-to-one correspondence between emotion categories and any pattern of neurobiological responses.
> $NOC_{Physiological}$: There is no one-to-one correspondence between emotion categories and any pattern of physiological responses.

$NOC_{Behavioral}$: There is no one-to-one correspondence between emotion categories and any pattern of behavioral responses.
$NOC_{Expressive}$: There is no one-to-one correspondence between emotion categories and any pattern of expressive responses.
$NOC_{Phenomenological}$: There is no one-to-one correspondence between emotion categories and any pattern of phenomenological responses.

By dividing NOC into subcomponents, we can have a better idea of what sources of empirical evidence would be relevant to test VT. However, this raises the question: is NOC, as a general claim, a conjunction (NOC_{Neural} & $NOC_{Physiological}$ & $NOC_{Behavioral}$ & $NOC_{Expressive}$ & $NOC_{Phenomenological}$) or a disjunction (NOC_{Neural} ∨ $NOC_{Physiological}$ ∨ $NOC_{Behavioral}$ ∨ $NOC_{Expressive}$ ∨ $NOC_{Phenomenological}$) of these sub-theses?

Interpreting NOC as a conjunction leads to an overly simplified claim. As soon as we find correspondence in one domain, NOC will be false. For example, if an emotion fails to correspond to neural, physiological, expressive, and phenomenological sets of responses, but corresponds to one common behavioral pattern, NOC is falsified. Consequently, we would have to reject VT, given that there is at least one domain where correspondence holds. This is a consequence that defenders of VT would find unacceptable, given that an important degree of variation would still hold.

On the other hand, interpreting NOC as a disjunction of the different sub-theses leads to an overly demanding claim to reject. In this case, evidence for lack of correspondence in one domain suffices to establish NOC and therefore to accept VT. Given that the aforementioned characterization of NOC includes domains where variability is expected (for example, in terms of action tendencies or expressions), rejecting NOC becomes not only implausible, but trivial. Evidence for some degree of variability in some domain abounds, rendering the question of variability almost insignificant.

In order to escape these problems, researchers must decide which domains offer the most relevant support for NOC. For example, basic emotion theorists would presumably hold the neural and physiological domains as more relevant than the phenomenological domain, given their commitment to the idea that emotions must correspond to patterns of physiological and neurological responses (see Ekman 1992; Ekman and Cordaro 2011; Izard 2007, 2009; Panksepp 1998, 2008, 2011; Panksepp and Watt 2011). Unfortunately, emotion researchers have not reached a consensus regarding the relevance of evidence in these domains. Without such consensus, it is impossible to judge whether empirical evidence supports NOC, hence precluding us from drawing any conclusions from this thesis.

Let us now turn to LC. As I explained above, LC is interpreted as a claim about the presence or absence of correlations between different measurements. Yet, there are three worries one can raise about this construal. First, given its appeal to correlations, it is unclear which correlations (or lack thereof) are necessary or sufficient to reject (or accept) LC. On the surface, we would consider the correlations between variables in each of the aforementioned domains. However, each of these domains counts with more than one variable. Consider the physiological domain. Among physiological measures used to study emotions, we find three families (cardiovascular, respiratory, and electrodermal), each with a wide range of possible measurements. Given the different possible variables researchers could employ in their studies, we can ask: do

we require correlations between all of these variables in all of these domains in order to reject LC? If not, which correlations suffice? And to which degree?

Besides this worry, a second problem with the current construal of LC is that it is unclear what it is for an emotion to obtain without anything we might call a package of responses. On a naïve understanding of the claim, this would be a case where an emotion obtains but no responses are observed. But then, why would we accept that an emotion obtains? If there are no neurological, physiological, behavioral, expressive, or phenomenological responses, there is no emotion either. It seems clear that this is not the intended interpretation. But which one is it then?

Perhaps the most plausible interpretation of the claim is that LC is true when an emotion obtains along with some responses, but these responses fail to correlate between each other. One such case would be an instance of fear where, for example, the neural and the behavioral responses fail to correlate. This may be due to there being several behavioral responses (fight or flight) with one common neural underlying response (for the sake of argument, suppose that this is amygdala activity; I shall go back to the nuances of this example below). Such an interpretation would also make the matter trivial though. For any emotion, there could be a myriad of possible behavioral patterns even if each emotion mapped one-to-one onto neural responses. The same may apply for expressive patterns, which vary depending on context (Elfenbein and Ambady 2002; Gendron et al. 2014) and do not map one-to-one onto many emotions. On this interpretation, LC becomes true by any sort of variation in any of these domains, making it a vacuous claim.

Someone might object that what we require then are criteria to individuate the patterns of responses referred to by LC. To avoid rendering LC trivial, the objection holds, we just need to specify on which level of abstraction we would consider a group of responses a pattern, so as to proceed to test correlations between these patterns and emotion categories. This move, however, raises a third worry that deserves special attention, namely, that it is unclear how we should individuate patterns of responses. In each of the domains in question, there are a number of ways in which researchers might consider a set of responses a pattern. This makes it difficult to decide the matter empirically. Without an answer to how to individuate patterns of responses, we cannot determine whether these patterns are correlated, hence leaving LC undetermined. Furthermore, this worry also affects NOC, since without criteria of pattern individuation, we cannot determine whether emotion categories correspond to a given pattern or not. In what follows, I will explore this difficulty in detail.

3 Individuating Sets of Responses

3.1 Neural Patterns

Traditional accounts of emotion that emphasized the role of neural mechanisms in emotions thought of emotions as relating to the activity in specific and consistent regions in the brain. According to these views, there must be something in the brain that is domain-specific to each emotion category. These are for example LeDoux's studies on fear conditioning (LeDoux 2003, 2007, 2013; see also Phelps and LeDoux 2005), which attempted to map fear onto amygdala activity, or Panksepp's (1998, 2011)

attempt to individuate subcortical structures underlying primary emotional processes. Following Lindquist et al. (2012), let's call these *locationist* accounts:

> [Locationist accounts] hypothesize that all mental states belonging to the same emotion category (e.g., fear) are produced by activity that is consistently and specifically associated with an architecturally defined brain locale […] or anatomically defined networks of locales that are inherited and shared with other mammalian species. (pp. 122-123)

We can distinguish two types of locationism. On one type, *anatomical locationism*, mental states belonging to the same emotion category correspond consistently and specifically to an architecturally defined brain region.[2] On a second type, *homological loactionism*, they correspond to inherited networks shared with other mammalian species.

According to anatomical locationism, Neural holds that a given emotion category corresponds to activity in more than one region or no specific region at all. In turn, LC holds that activity in a specific brain region does not correlate with responses in other domains. Empirical evidence supports Neural in the anatomical locationist sense. Early meta-analyses showed some promising results mapping, for instance, fear to the amygdala or sadness to the subcallosal cingulate cortex (Murphy et al. 2003; Phan et al. 2002). However, they also suggested some degree of overlap. For example, the fact that both happiness and disgust were associated with basal ganglia activation suggests that these areas are not unique to either emotion, but rather are involved in some general process.

Later meta-analyses stressed this kind of findings. Most famously, Lindquist et al. (2012) analyzed a considerable amount of studies, including those analyzed in previous meta-analyses, claiming that correspondence between emotion categories and brain locations failed to obtain. For example, they claimed that the amygdala, a region traditionally associated with fear, was instead involved in "signaling whether exteroceptive sensory information is motivationally salient" (Lindquist et al. 2012, p. 130), since it had also been observed in other tasks such as orienting responses to motivationally relevant stimuli, novel and unusual stimuli. Moreover, lesions to the amygdala do not only affect fear responses, but also responses to other relevant stimuli in general. Additionally, their analyses revealed that amygdala activity was also significantly associated with disgust, indicating some degree of overlap between different emotions. Similar claims followed for the rest of the so-called basic emotions. For every candidate region that would correspond to a given emotion category, Lindquist et al. argued that it was involved in more general processes and hence that there was no correspondence between emotion categories and the activity of specific brain regions.

The aforementioned results offer good reasons to accept NOC_{Neural} provided we adopt anatomical locationism. But what about adopting homological locationism instead? According to homological locationism, the relevant level at which we should individuate neural patterns is not at the level of specific anatomical locations, but at the level of networks that are inherited, anatomically constrained, and have homologues in other mammals (Panksepp 2008). Thus, NOC_{Neural} would not be established by failure

[2] Scarantino (2012a) calls this *radical locationism*.

to find corresponding activity in specific brain locations, but rather if we fail to find innate, anatomically intrinsic networks.

One example of a homological locationist view is Panksepp's (1998, 2011). Panksepp claims that in order to individuate the neural patterns in the brain, we must rely on comparative studies using non-human animals to find evolutionarily adapted networks. He presents evidence for subcortical networks that interconnect midbrain circuits with various structures in the basal ganglia, such as the amygdala and the nucleus accumbens, through pathways running through the hypothalamus and thalamus. Among the networks he identifies, he includes SEEKING, RAGE, FEAR, LUST, CARE, PANIC, and PLAY networks (Panksepp 2011).

Evidence for homological locationism often involves studies in non-human animals. Studies of this sort include studies on fear conditioning (LeDoux 2003) or studies on subcortical circuits (Panksepp 1998, 2011). Research of this sort is still promising and suggests that there may be inherent networks underlying at least some emotional reactions. For example, Yilmaz and Meister found that mice reliably engaged in escaping or freezing behavior when a specific visual stimulus was presented. More interestingly, the researchers could manipulate the probability that mice would engage in either of these behaviors depending on the physical properties of the stimulus, suggesting that these physical properties activate automatic, rapid firing inherent circuits underlying fear reactions (Yilmaz and Meister 2013).

There is nevertheless evidence against the presence of intrinsic networks in the brain as well. Touroutoglou et al. (2014) show that increases in activity during emotion experience and perception do not map onto intrinsic networks in the brain using resting state connectivity fMRI. Instead, they report finding domain-general networks involved in emotional experience, which in their view supports the claim that there is nothing we can call a coordinated package of neural responses in terms of intrinsic networks. For example, for fear, sadness, and happiness, they found a general dorsal region connecting the anterior insula and the anterior cingulate cortex. Consequently, they conclude that even adopting homological locationism, Neural would still be established.

Besides locationism as a general view of emotion-brain mapping, a second view has emerged in recent years, one stemming from a functionalist framework. Instead of trying to individuate patterns in terms of domain-specific intrinsic networks, we could take the relevant neural patterns to be at the level of regions that show correlated activation even in the absence of an intrinsic network, i.e. functional networks.

A prime example of this approach is the one involved in multivariate pattern analyses (MPVA). One of the earlier studies using these techniques is Kassam et al. (2013). In this experiment, subjects saw different emotion words while inside a MRI scanner and were asked to attain the corresponding emotional state for a period of time. The researchers then trained a classifier on fMRI data in order to test whether the classifier could accurately predict the subject's emotional state. Kassam et al. report that their classifier was able to successfully predict a subject's emotional state from their neural data in a given trial. They report that this classification was accurate between 77% and 89% of the time. After Kassam et al. (2013), other studies followed suit. Using film and instrumental music induction followed by self-report, Kragel and LaBar (2015) managed to classify seven emotional states using MVPA on neural activation

data (contentment, amusement, surprise, fear, anger, sadness, and neutral) with 37.3% accuracy (chance = 14.3%).

One recent study has gained special attention among defenders of multivariate approaches to emotion: Saarimäki et al. (2018). In this study, participants heard 4 narratives for each of 14 emotional states plus a neutral condition. The classifier in this case managed to classify most of the target emotions. Altogether, 12 emotions (excluding longing and shame) could be reliably classified from fMRI signals. From these findings, Saarimäki et al. conclude that multiple emotion states have distinct and distributed neural bases. In their view, many emotions are represented in the brain in distinct yet overlapping regions. They claim that each emotion state likely modulates different patterns measured with fMRI, and the overall configuration of the regional activation patterns defines the resulting emotion.

As explained above, it is crucial to these findings that the classification of brain activity corresponding to an emotion is not in terms of specific regions of activation, but as patterns throughout the whole brain. This substantially affects questions about coordination and correspondence. In the first case, we must decide at which point a pattern counts as delimited enough to be considered a candidate to correspondence. In the second, we must decide whether correspondence with networks counts as explanatorily relevant in the context of variability. In other words, we must decide whether mapping emotions onto brain networks is explanatorily interesting or whether we must keep the level of analysis at the level of locations.

Questions of this sort are reminiscent to debates on *cognitive ontology* (Anderson 2015; Price and Friston 2005). A cognitive ontology consists of criteria to map cognitive functions to anatomical structures, such that we can infer one from the other. This includes characterizing what the specific functions of particular brain structures are and how these functions contribute to the overall function being addressed by a given cognitive task. The case of emotions is analogous: we must have a set of criteria to map emotions onto brain structures. Furthermore, just as in the case of cognitive function, this may call for different approaches. Anderson (2015) distinguishes three views, depending on the degree to which researchers must revise their current psychological framework. We can either attempt to preserve as much of the current framework as possible, revise it until it fits our best characterization of brain activity, or modify it radically to the point of revising our theoretical primitives themselves. I will not attempt to defend a particular approach here, since this requires a detailed discussion of both theoretical and empirical claims. Nevertheless, it is a decision that is pending in emotion research, one that only until recently has begun to be put on the table (Celeghin et al. 2017; Scarantino 2012b).

3.2 Physiological Patterns

Besides looking into neural activity associated with emotion, another important source of evidence comes from studies on physiology. Physiological activity in the context of emotions generally refers to three types of autonomic[3] variables. First, there is activity

[3] Strictly speaking, not all physiological responses concern the autonomic nervous system. Other physiological responses include, for example, muscle tension, or even neural responses. In order to avoid cashing out neural responses as a subset of physiological responses while making clear what these refer to, I shall use the term physiology as interchangeable with autonomic.

related to the cardiac system, which includes heart rate variability, blood pressure, cardiac cycles, and the like. Second, we find variables regarding respiration, e.g. respiratory cycles, respiration period, amplitude, etc. Lastly, there are variables concerning electrodermal activity, i.e. skin conductance levels, responses, resistance, etc.

Given these types of physiological variables, whether or not there are coordinated patterns of physiological activity can be broken down into two criteria. One is determining whether there is patterning within a class of variables. We can ask whether there are specific patterns concerning cardiac, respiratory, or electrodermal activity for a particular emotion. Additionally, we can investigate patterning between the classes of variables, i.e. whether cardiac, respiratory, and electrodermal activity are robustly correlated and form a homogeneous set of responses for each emotion.

Early studies of physiological activity tried to show autonomic constants across different emotions. In one of their first studies, Ekman et al. (1983) asked subjects to contract specific muscles in order to mirror implicitly a given facial expression without telling them which expression it was, or asked them to relive an experience that would elicit a given emotion. During these tasks, the investigators measured the subjects' heart rate, left- and right-hand temperatures, skin resistance, and forearm muscle tension. Ekman et al. report that autonomic variables change significantly depending on the emotion. They found that heart rate and temperature increased for anger, as well as heart rate increases for fear, in contrast to happiness. The researchers also hold that they were able to distinguish disgust from anger, fear, and sadness in the first task, and sadness from disgust, anger, and fear in the second. In a later report (Levenson et al. 1990), the same researchers report similar findings.

As in the case of neural activity, more recent studies have introduced multivariate techniques to look for physiological patterns. Rainville et al. (2006), for example, used principal component analysis (PCA) to see which variables were the most useful when classifying different emotions from data on autonomic responses. As in other studies, they used autobiographical recall methods to elicit anger, fear, happiness, and sadness. They measured a number of variables regarding respiration and cardiac cycles, including respiration period, amplitude, heart-rate variability, and others.

In their analyses, the researchers report some differences in these variables as a function of emotion. For example, they claim that respiratory period decreased in fear and happiness and less consistently in anger, while the variability in respiratory period increased in sadness. Overall, the researchers claim that this study provides some evidence that basic emotions are associated with distinctive patterns of cardiorespiratory activity. Different emotions were distinguished from a neutral condition based on different subsets of dependent variables and multi-dimensional exploration of the data revealed complex patterns of activity that characterized each emotion. According to the PCA, the variance in cardiorespiratory activity can be explained along five dimensions, mostly related to heart-rate variability.

Recent meta-analyses also reveal some autonomic specificity for emotion. Kreibig (2010), for instance, covered 134 publications and examined three classes of variables: cardiovascular, respiratory, and electrodermal. She reports specific patterns for a great number of emotions. For example, she claims that anger involves faster breathing as seen in shortened inspiration and expiration times, more expiration than inspiration, increased heart rate, increased overall blood pressure, among others. Fear elicited a

similar pattern, involving broad sympathetic activation, cardiac acceleration, increased vaso-constriction, and increased electrodermal activity. However, in the case of fear, peripheral resistance decreased whereas it increased for anger.

Despite the studies supporting autonomic specificity, there are also important challenges. In a recent meta-analysis, Siegel et al. (2018) evaluated empirical evidence in favor or against specificity. The authors included 204 studies from 1950 to 2013, and studied the same three-fold division of variables used by Kreibig, namely, cardiovascular, respiratory, and electrodermal measures. To do this, they compared the effect sizes across all their studies. Some results display mean ANS changes from baseline across several effect sizes but with substantial variability. For instance, the patterns of anger and fear showed large effect sizes, suggesting that their physiological patterns differed significantly from baseline across several autonomic variables (specifically, heart rate, cardiac output, diastolic and systolic blood pressure). Yet, these effect sizes are very heterogeneous, indicating that even though these emotions have clear physiological effects, these effects are not uniform and do not form a stable pattern.[4]

Other results show small mean ANS changes and moderate variability. For example, the researchers report the cases of disgust and neutral categories. For disgust, only skin conductance level and responses had relevant effect sizes, but only the latter was homogeneous. For neutral conditions, only systolic blood pressure had an interesting mean effect size, but it is also a heterogeneous variable. Happiness and sadness had increased effect sizes in heart rate, diastolic blood pressure, skin conductance level, and others, but they are mostly heterogeneous. As a result, most mean ANS changes were not uniform. Additionally, the researchers claim that ANS changes were not specific to a given emotion category either. Happiness had a mean increase in skin conductance level similar to disgust, anger, fear, and sadness, for example.

We can now interpret evidence challenging physiological specificity using the categories presented above. On one hand, there is evidence suggesting that there is low coordination between cardiac, respiratory, and electrodermal variables. Evidence of the first type presented by Siegel et al. is one example. As they suggest, correlations between physiological variables preclude their classification as a specific pattern. On the other hand, there seems to be evidence showing low coordination within physiological variables, as presented in the second group of findings reported by Siegel et al. According to this argument, some physiological variables have more impact than others in determining the ensuing emotion. As a result, given the lack of correlation between physiological variables, the researchers claim that there is no physiological specificity for emotion.

Nevertheless, settling this discussion requires further methodological and epistemological decisions. First, it is unclear whether all physiological variables should have the same influence when considering whether there is a coordinated pattern or not. Often used variables such as heart-rate variability surely are among the most important ones to consider. Yet, the status of other variables such as respiration period or vaso-constriction is left undecided.

[4] It is worth noting that heterogeneity in effect sizes might be accounted for by differences in intensity rather than physiological variability. This would mean that variability more of an artifact rather than a metaphysical fact about emotions themselves.

Second, both optimistic and skeptical researchers fail to distinguish between within- and between-variable coordination. This leads to an ambiguity that affects both camps. On one hand, it could be the case that we need correlations among variables of the same type (say, respiratory variables) in order to consider that there is a robust physiological pattern (i.e., a respiratory pattern). On the other hand, we may not demand correlations within a given family of variables, but rather between some measures of different types. For instance, we may expect some cardiovascular measures to correlate with some electrodermal ones, without the requirement that there are within-variable patterns. As it stands now, researchers highlight evidence showing that one measure is associated with a given emotion or that another is not, without a clear argument as to whether it is necessary that all measures of a given type correlate with one another or whether it is necessary that some measures of different types do so.

Lastly, similar to the discussion regarding neural patterns, the use of multivariate techniques is still controversial. Presumably, lots of physiological processes obtain when we experience an emotion or any other state. As a result, whether or not the ability to classify them with analyses such as PCA tell us something explanatorily relevant remains unclear. Skeptics may argue that the mere presence of a statistical pattern says little about the causal mechanisms involved in emotion. Optimists may react by pointing out that multivariate techniques are nevertheless more robust and that they do not claim that there is just any pattern at play.

3.3 Behavioral Patterns

Behavioral patterns refer to possible behavioral outcomes of an emotion episode. In the current literature, the best account of the behavioral patterns of emotion comes from appraisal theories. According to these theories, emotions involve states of *action readiness* (Frijda et al. 1989). On one influential construal, action readiness is cashed out as the individual's readiness or unreadiness to engage in interaction with the environment. This may consist in readiness to engage or disengage from interaction with some object in a particular way (*action tendency*) or in a general state of activation or inhibition of behavior (*activation modes*) (Frijda 2007).

Some researchers claim that we can differentiate between emotions by appealing to the different states of action readiness they elicit. On one such study, Frijda et al. (1989) asked subjects to recall instances of emotions and asked them to rate different statements concerning various action patterns. These statements included descriptions such as "I wanted to approach or make contact" or "I wanted to oppose, to assault." They then tried to map patterns of action to emotion names by investigating how well they could predict the emotion label from these patterns. Frijda and colleagues report some predictability for 32 emotion categories. Among the highly correlated patterns they report crying for sadness, protecting one self for fear and anxiety, moving against an object for anger, avoidance for disgust, and hiding from others for shame, among others. This suggests that there may be some correspondence between action readiness states and emotion categories, speaking against $NOC_{Behavioral}$.

Other studies have yielded similar results. Roseman et al. (1994) used as similar experimental design, asking subjects to recall past emotional experiences and answer a questionnaire that tapped into their behavioral outcomes. The researchers claim that their experiment shows clear distinctions between 12 emotions in terms of their action

tendencies. Among the tendencies reported, we find fear as a readiness to reduce the possibility of harm, sadness as crying and seeking comfort, disgust as attempting to get something noxious out of the body, among others.

In spite of these optimistic efforts, cashing out emotions in terms of action readiness and action tendencies does not go without problems. On one hand, there is some observed variability. In Frijda et al. (1989) we find one action pattern corresponding to two emotions (e.g. protecting oneself in fear and anxiety). On the surface, this would only mean that variability in terms of behavior is still controversial. However, the problem runs even deeper.

Critics of appraisal theories have argued that the links between emotions and action tendencies may as well be a matter of conceptual truth rather than empirical fact. If so, questions about correspondence become trivial; emotions will trivially correspond to behavior patterns (just as water corresponds to H_2O.) Consider the presumed correspondence between fear and engaging in behavior towards protecting oneself in situations of perceived harm. Suppose we attempt to falsify such correspondence. We would need to be able to obtain a fear state that does not involve such a behavioral tendency. Yet, arguably, that tendency is precisely what it means to be in a fearful state. As a result, any candidate state to falsify this supposed hypothesis would not count as a fear state as a matter of conceptual fact.

This problem can be brought to light by considering moves to ameliorate it. Roseman (2011), in response to the variability between emotions and behavioral outcomes, argues that emotions are consistent at the level of coping strategies. He claims, for instance, that fear forms a consistent pattern insofar as it involves a strategy to quickly and urgently move away from or stop moving toward some danger (or at least some description of the sort). Even if we sophisticate such a description to involve other behavioral aspects of fear, we could still ask: what would it mean for this description to be inadequate (i.e. for its correspondence with an emotion category not to be the case)? Presumably, this correlation obtains, not as a contingent fact, but because the behavioral outcome provides a definition of what it means to be afraid. To use Smedlund's (1992) example, these results are as if we discovered that bachelors are male and single.

A similar worry runs regarding LC. Any behavioral outcome that may correspond to a give emotion, even if not one-to-one, can be spelled out to yield a correlation with some neural and physiological state. In this sense, coordination between the neural and physiological domain would be almost trivially true. If we cut out behavioral patterns with too fine a grain, we risk rendering these correlations meaningless. One can resist this result by clarifying that triviality only obtains if neural and physiological states are interpreted as token states, not as types, i.e., by adopting a coarser grain. Still, the question of how to spell out these patterns properly remains unanswered.

3.4 Expressive Patterns

The issue of whether there are expressive patterns for each emotion has its roots in the issue of the universality of facial expressions. Broadly construed, the question is whether there is a set of universally recognized and produced expressions corresponding to each emotion (or a subset thereof). As Russell (1994) formulates it, the thesis of universality can be divided into four propositions:

(a) Specific patterns of facial muscle movement occur in all human beings.
(b) Certain facial patterns are manifestations of the same emotions in all human beings.
(c) Observers everywhere attribute the same emotional meaning to those facial patterns.
(d) Observers are correct in the emotions they (consensually) attribute to those facial patterns. (cf. Russell 1994, p. 106)

Let us focus on the first two propositions.[5] The first of these propositions relates to the specificity of the facial expressions themselves (i.e. to their coordination). The second, to their correspondence to emotional states (i.e. $NOC_{Expressive}$).

The main defender of universality is Ekman (see e.g. Ekman 1972, 1980; Ekman and Friesen 1971; Ekman et al. 1969, 1983, 1987). Ekman has conducted a number of studies allegedly establishing the universality of at least some facial expressions (those corresponding to the so-called basic emotions). Among these studies, perhaps they most often cited is that by Ekman and Friesen (1971). Ekman and Friesen (1971) conducted a study in a remote culture that would have had no contact with Western cultures, the Fore group in New Guinea. In the experiment, the researchers showed subjects three photographs of different facial expressions and told them a story. Subjects then chose the photograph that matched the story's emotional content. Stories included content for happiness, sadness, anger, surprise, disgust, and fear. Ekman and Friesen report that subjects were generally able to identify the correct photograph at a high success rate. In their view, this result provided evidence that there were facial expressions that were universally recognizable, even in cultures with no contact with Western societies. In later years, Ekman et al. (1987) would use this method in other Western cultures and replicate these findings.

Even though these findings are sometimes taken as granted, universality is still controversial. Arguments against universality come in two main strands. The first strand of criticism intends to cast doubt on the robustness of the findings presumably supporting universality, showing flaws in the designs as well as the assumptions of a number of studies. The second strand tries to outweigh empirical evidence for universality by underscoring cultural variation. Whereas the first strand presupposes that we already know what evidence is relevant for the question of universality (it only claims that researchers have failed to obtain such evidence), the second strand concerns precisely what type of evidence would establish universality or not, thereby offering grounds for or against $NOC_{Expressive}$ and LC. Consequently, I shall focus on the second strand.

Attacks of this type on universality come from two sources. One is evidence showing that agreement among cultures regarding which facial movements correspond to which emotions has been overstated. Rather than finding robust agreement, researchers have showed that agreement drops under certain conditions. One example is the meta-analysis by Elfenbein and Ambady (2002). Using the same data as that in Ekman's studies, among others, Elfenbein and Ambady show that members of a given group are more accurate in judging expressions of members of their same group.

[5] The third and fourth propositions relate to the observers of these facial expressions. Given that these propositions concern our attribution of emotions through expressions rather than the presence of the expression themselves, I will leave them aside.

Specifically, they report that Western participants are 9.3% more accurate when judging other Western faces than with African or Asian ones. Even if overall recognition is still above chance level, these results suggest that accuracy scores depend on culture and are not as uniform as defenders of universality might think.

The other source of evidence against universality are studies showing differences between different populations in terms of their perception and categorization of facial expressions. Examples involve studies in a number of cultures. Elfenbein and Ambady themselves claim that when analyzing data in terms of individual emotions, some emotions are poorly recognized universally. In their meta-analysis, they found that fear and disgust are the most poorly recognized, even though they are among the most cited candidates to universal, biologically basic emotions. According to them, this implies that culture still shapes meaning of faces even in the presence of some uniformity.

Other studies also show mismatch between Western and non-Western interpretations of faces. Crivelli and Fridlund (2018) report that communities from the Trobiand islands in New Guinea understand gasping faces as threat displays instead of fear displays, as traditional studies have attempted to show. Similarly, Gendron et al. (2014) found that the Himba people of Namibia perceive facial actions in context, that is, not as corresponding to a feeling but to the whole situation. For example, instead of interpreting crying as corresponding to a feeling of sadness, they situate it as a response to death. Jack et al. (2012) report that East Asian facial expressions overlap, leading to fuzzy categorization contrasting with Western taxonomies. Along the same lines, Jack et al. (2016) claim that in Chinese societies categorize emotions into more categories than English samples when asked to judge facial expressions.

Evidence for cultural variation in emotion expression production and recognition tries to dismantle the second proposition presented above, namely, that emotional expressions are manifestations of the same emotions in all humans. On the face of it, there could be universal patterns of expression (assuming the methodological criticism is misguided), but they do not correspond to the same emotions everywhere. If this is true, then we would have evidence to reject LC, but also we would have to accept $NOC_{Expressive}$.

Apart from showing how universality is controversial, I suspect these findings suggest that the question of the universality of emotional expression is a different topic altogether that does not have much bearing on the issue of variability. Even if there were no universal patterns of emotional expression (LC) and hence no correspondence with emotions ($NOC_{Expressive}$), would that entail that emotions are variable phenomena? Plausibly not. The reason is that there could still be fixed, even innate patterns at the neural or physiological level that would grant emotions some robust form of homogeneity. Otherwise, we would be forced to split emotion categories in terms of their different expressions even in presence of evidence for neural and physiological homogeneity. If this line of argument is correct, it follows that evidence on the universality of emotional expression is at best unnecessary to determine whether variability is the case or not. Even if universality of expression wasn't the case, there could still be good reasons—perhaps even stronger reasons—to decide for or against variability, such as the potential presence (or absence) of homogeneous physiological or neural patterns.

Moreover, I suggest that not only is universality unnecessary to decide variability, but that it is also plausibly insufficient. Let us assume that universality of expression is the case. Let us also assume, for the sake of argument, that there is no homogeneity at other levels. In such a case, we would have a mapping from one fixed emotional

expression to many neural, physiological, behavioral and phenomenological patterns. This quite plausibly would not qualify as evidence to reject variability, as there would still be solid grounds to hold emotions as naturally disjoined phenomena even if they map onto specific expressions. The absence of homogeneous patterns at all other levels would indicate that emotions lack the sort of unity that might be relevant to form the types of kinds a science of emotions looks after. While it is true that this picture is empirically implausible, given that universality of expression would probably entail other kinds of patterns (e.g. fixed physiological patterns), this suggests that universality is perhaps only interesting as a proxy for other types of patterns rather than a way to decide variability by itself.

Additionally, someone may object that expressions are still a central part of our emotion attribution and behavioral manifestation. Yet, this would then reduce expressions to subsets of behavioral patterns. As a consequence, there would be no reason to consider expressive patterns as separate from behavioral patterns (action tendencies) altogether. In other words, expressions are at best part of our action tendencies, and at worst irrelevant to the case of variability.

3.5 Phenomenological Patterns

Phenomenological patterns are often understood as patterns of subjective experience. What exactly characterizes subjective experience is nevertheless unclear. Subjective experience, taken as a criterion to individuate patterns candidate to correspondence and coordination, fails on two grounds. First, the phenomena subjects and theorists describe as subjective can plausibly be reduced to other types of patterns already under consideration, such as collections of neural and physiological states as well as action tendencies (behavioral patterns). Second, even if there is some remainder in terms of qualitative experiences, there are good reasons to doubt these can successfully help us individuate emotions, thus precluding claims even about their variability.

A first approximation to tap into phenomenological patterns of emotions is to rely on self-report data, asking subjects to narrate or describe their own emotional experience. One influential example of such an approach is the work by Davitz (1969). Davitz undertook to develop a dictionary of emotions that synthesized how people use language to refer to their own emotional states. Based on a short list of emotion terms (Affection, Anger, Anxiety, Boredom, Cheerfulness, Confidence, Impatience, Sadness, and Satisfaction), he interviewed people asking them how they would describe each state and recorded their reports. To this list of statements he then added more descriptions from 1200 subjects who were asked to think of concrete instances of each emotion. From these reports Davitz obtained a list of 556 statements about emotion experience. Lastly, he asked a third group to rate how adequate each statement was to describe their own experiences. With this material in hand, he compiled the most used descriptions for each term into the dictionary.

A short examination of the definitions and statements found in Davitz's dictionary shows that many of the descriptions presumed to tap into subjective experience can be reduced to other patterns. For example, 'anger' includes among its most common descriptions 'my blood pressure goes up,' 'my pulse quickens,' or 'my heart pounds.' In the case of sadness, we find 'there is a lump in my throat,' 'there is a clutching, sinking feeling in the middle of my chest,' and 'I have no appetite.' Similar descriptions

can be found for other emotions terms as well. In these cases, it is easy to see that these patterns can be described as physiological patterns corresponding to each emotion. Other patterns in Davitz's dictionary's entries refer rather to action tendencies. In the cases above, anger includes 'my fists are clenched,' 'there is an impulse to hurt, to hit, or to kick someone else,' and sadness, 'tears well up' or 'I cry.'

Perhaps a more sophisticated approach to the phenomenology of emotion is found in Lambie and Marcel (2002). Lambie and Marcel distinguish three types of questions regarding the empirical investigation of emotions: (a) what is the content of emotion experience as it is experienced?, (b) to what nonconscious process or representation does emotion experience correspond?; and (c) what processes or differences in content lead to and contribute to emotion experience? In their view, only the first of these questions tackles phenomenology.

To answer the question of what is the content of emotion experience, Lambie and Marcel separate between emotion states and emotion experiences. Emotion states are the functional aspects of emotion apart from conscious experience, which include primary appraisals of events in terms of relevance to the organism, the activation of brain and bodily systems, and preparation for action. Emotion experiences are both the phenomenological aspects of emotional states (first-order experience) and the awareness of these experiences themselves (second-order experience).

For the purposes of evaluating NOC and LC, following Lambie and Marcel, we would have to decide at which level are emotions individuated. Suppose we decide that they must be individuated at the first-order experience level. Characterizing first-order experience is problematic for a number of reasons. As Lambie and Marcel recognize, our first mode of access to first-order experience is by introspection, which requires awareness of it, which in turns changes the first-order experience itself. The authors suggest that we can instead rely on memory and episodic reinstatement to tap into previous episodes of first-order experience, circumventing this problem. Yet, there is good evidence on memory manipulation showing that memory is also affected by our current epistemic states (Brown and Marsh 2008; Edelson et al. 2011; Loftus 2005; Mazzoni and Memon 2003). If this is the case, relying on episodic reinstatement does not fix the problem.

A second worry regarding first-order experience is that, as Lambie and Marcel have characterized it, it includes aspects that are again reducible to other patterns. Given that first-order experience has underlying brain and bodily states, as well as involving action tendencies and appraisals, it is unclear why this level of description would yield a different type of pattern at all. As the characterization stands, Lambie and Marcel have suggested facets of neural, physiological, and behavioral patterns that are involved in emotion, but have not shown that there is something uniquely phenomenological worth separating.

One may object that these reductions leave the qualitative character of emotion experience untouched. In the same vein as proponents of the explanatory gap in philosophy of mind (Chalmers 1997; Levine 1983) one could argue that physical or behavioral states do not exhaustively describe pure forms of emotion experience. An argument of this sort seems to be in the background of LeDoux's (2012, 2013; LeDoux and Brown 2017) claim that emotions and feelings should be used interchangeably. As a consequence, LeDoux recommends not making reference to emotions when we talk about circuits underlying survival behavioral dispositions, and instead looking for a theory of emotional consciousness as a theory of emotion.

This approach would entail individuating emotions by their qualitative character alone. As the record of discussions on the hard problem of consciousness attests, problems soon arise. First, to do this we need an account of how we could ground emotion concepts on first-person qualitative properties. Since we presumably do not have access to others' first-person experiences, we seem to fall prey of arguments such as Wittgenstein's private language argument (Wittgenstein 1953/2009). According to a broad and naïve construal of this argument, it is nonsensical to think that the meaning of concepts such as 'red' or 'yellow' (and in this case 'sadness' and 'fear') can grounded in first-person experience, since we would have no public criteria for their correct application. This leaves us with concepts on which we cannot construct a scientific theory.

Second, we could resist the private language argument (and other similar ones) and insist that there is no reason why it would be impossible to ground phenomenal properties on publicly available criteria. There are good reasons to doubt that reductions of the phenomenal character of consciousness are impossible in principle (see e.g. Pauen 2017), hence opening the door for third-person descriptions. In other words, we can reject the explanatory gap and defend the possibility of describing phenomenality in functional terms. This is for instance what LeDoux and Brown (2017) attempt in offering a theory of emotional consciousness. If this were so however, then we would be able to describe phenomenality in neural, physiological, or behavioral terms, thus diluting the category of phenomenological patterns into the other three.

Lastly, it is doubtful that the appeal to irreducible qualitative properties provides a tractable account at all. It is difficult to see, on their qualitative aspects alone, how emotions can differ from one another. This case is clear for emotions that are similar to one another like anger and indignation, or joy from pride (Prinz 2007, p. 52). However, it is even more pressing for comparisons between intuitively very different emotions such as anger and fear, which are more similar to each other than, for instance, happiness and fear. Without invoking non-qualitative properties such as valence (which is ultimately a relational property),[6] we cannot explain many differences between emotions (for an argument in this direction, see Frijda et al. 1989, p. 227).

Let us grant then that first-order experience cannot do the trick. We may still claim that phenomenological pattern individuation can obtain at the level of second-order awareness. Second-order awareness can be characterized, according to Lambie and Marcel, in two ways. In some cases, our emotion experience is directed to the self. These are cases where we, for example, experience anger as an offense to our own selves or sadness as an own failure. In other cases, emotion experience is directed to the world. Here our emotions are describable in terms of objects, as when we experience

[6] I am grateful to an anonymous reviewer who pointed out that qualitative properties might be relational. They consider the case of orange differing from yellow in that it is more red. In my view, while it is true that we might be able to describe some qualitative properties relationally, we must also assume a fixed base on which to ground such descriptions. In the case of orange, it is because we have a fixed set of primary colors which include yellow and red that we can describe other colors relationally. In the case of emotion however, it is unclear which qualitative base we can assume to ground such descriptions. The reason why valence does not provide a good candidate is that since it is relational, it must be defined in virtue of other non-relational qualitative properties for which there is no good candidate and that even if there were, as I suggest before, they would fall prey of the private language argument. Hence, it is difficult to see how to ground descriptions of different emotion categories based on irreducible qualitative properties, which by extension would undermine the description of emotion categories in virtue of relationally described qualitative properties.

the object of our anger as something offensive or blameworthy, or sadness as presenting a world that is unfulfilling.

The case for phenomenological pattern individuation at the second-order awareness level resembles attempts in appraisal theories to individuate emotions in terms of *core-relational themes* (Lazarus 1991). It is also reminiscent of other approaches in phenomenology proposed by enactivists (Colombetti 2009, 2017; Hutto 2012). In both of these cases, second-order awareness refers to an experience of an emotion in terms of the relation between an organism and its environment, whether it is focused on the standing of the self as related to objects or focused towards properties of the objects as appraised by the self.

If this interpretation is correct, second-order awareness may be described in terms of other patterns as well, namely, as action tendencies. Both appraisal theorists and enactivists stress the idea that the phenomenology of emotion, so construed, is essentially linked to our possibilities of action given a relation with the environment. In a broad understanding of action tendencies, we can describe these relations as possible behavioral outcomes an organism may experience in a given moment. Again, phenomenology is described as part of other patterns already considered, casting doubts on the decision to separate it into its own category.

As a result, the individuation of phenomenological patterns as a separate category of patterns that would be candidates for correspondence and coordination seems unpromising. Either we get stuck with problems in grounding concepts in first-person experience, hence precluding us from establishing any claims regarding their variability, or, if we can overcome such an obstacle, we would be able to reduce phenomenal patterns to other patterns which turn out to be the relevant ones to decide for or against variability. Consequently, I propose leaving the qualitative character of emotional experience separate from the problem of variability or taking it as a result of other relevant patterns.

4 How to Evaluate Variability

In this article, I have argued that the Variability Thesis (VT), as construed in emotion research at the moment, is an ill-defined thesis. I raised some ambiguities regarding some of its logical properties, and then suggested that apart from these problems, there are difficulties spelling out what counts as a pattern to be candidate for correspondence and coordination. Without an answer to these issues, researchers cannot properly judge empirical evidence for or against this claim, leaving much of the debate in a stalemate.

Before closing, I would like to sketch some suggestions about how to understand VT in a scientifically fruitful way. It must be clear in advance however that a detailed discussion of the consequences of these desiderata requires more space than what is left here, but I shall nevertheless attempt to leave some important lessons on the table from which a science of emotion can benefit.

First, it is central to offering an account of variability that researchers become explicit about what evidence they consider relevant to split or lump together emotion categories. This includes questions such as whether neural or physiological differences suffice or not to split an emotion category, or whether behavioral homogeneity is enough to consider an emotion as a whole. Since different theories have different commitments in this regard, it is important to at least be explicit about them so that

researchers can evaluate variability in each theories' terms and detect when a theory fails to meet their empirical predictions while doing justice to their claims.

One important application of this consideration lies in the discussions between psychological constructionists and basic emotion theorists. Psychological constructionists have argued heavily for the claim that empirical evidence supports variability, while (at least some) basic emotion theorists have denied this claim. To decide these issues, it is vital that each theory makes clear on which conditions they hold variability to be true. As the current debate stands, it is unclear for example whether psychological constructionism is compatible with locationism (emotions could be potentially constructed even if they are constructed in specific regions in the brain, even though constructionists often reject locationism) or whether basic emotion theory is incompatible with variability (but see Scarantino (2012b, 2015) for arguments to the contrary). In any case, theories must make their commitments explicit in order to discuss empirical evidence on the same ground, otherwise risking talking about different claims altogether.

Second, as I have hinted at above, researchers should focus on neural, physiological, and behavioral evidence. This is because, on the one hand, it is unclear whether evidence on the universality of expressions can inform questions about variability, and if it can, it seems that it is reducible to behavioral patterns. On the other hand, regarding phenomenological patterns, a similar argument obtains, i.e., either they do not inform the issue of variability or they are arguably reducible to patterns of behavior.

By focusing on certain kinds of patterns, researchers can have a better grasp on which kinds of evidence are relevant to decide variability and its consequences. This would not only inform our consideration of previous evidence, such as leading to a reconsideration of the role of the universality of emotional expression and its alleged correspondence with specific physiological patterns, but would also serve to delimit and specify which future studies will be relevant to decide fundamental claims about the nature of emotions. In particular, I suspect that focusing on these three kinds of patterns will lead to a more detailed discussion on the merits and disadvantages of MVPA for the individuation of neural patterns, a clearer idea of what to expect at the physiological level, and more conceptual clarity on what behavioral patterns entail for the distinction between different emotion categories. In all three cases, we would have a better grasp of which empirical evidence is relevant and where to look for it.

This leads directly to a third desideratum: it is crucial that researchers agree on the criteria to individuate different patterns. This involves agreeing on the neural ontology of emotions (anatomical, homological, or functional locationism), the types of correlations that support a physiological pattern, and how behavioral patterns can be individuated without circularity and triviality. Without these criteria, researchers will keep on talking past each other without a consensus on how to evaluate the empirical evidence at hand. While this is difficult in practice, it is paramount that researchers pay attention to these debates in order to better understand the import of different kinds of empirical evidence. For example, should we find no specific region in the brain for a given emotion category (as constructionists often claim to have found), does that mean that there is no useful neural description that we can map onto that category? Would that also apply for the alleged lack of consistency at the physiological level? And what do results in behavioral psychology offer to the question of what emotions are? These questions, I take it, must be addressed in order for empirical evidence to come to have bearing on the more general questions emotion theorists are after.

Lastly, it is worth asking: if variability were truly the norm, by what criteria do we manage to apply emotion categories in everyday life? In other words, is there some relation between folk emotion categories and VT? At this point, it is difficult to say given the difficulties in understanding what exactly VT claims as an empirical hypothesis. Nevertheless, it is worth pointing out that we may still apply folk emotion categories even if VT turns out to be true. This would suggest some unity at a higher level of abstraction than what a purely physicalist framework—a framework that has been assumed for the most part in the debate so far—can offer. In this direction, researchers should consider the possibility that the unity of emotion categories lies, not in specific correspondences, but in more complex, multiply realizable systems.

Hopefully, these desiderata can help make clear what the points of debate are and how different theories relate to empirical findings. Even though many of these discussions require a deeper treatment than I was able to offer here, I expect to have raised questions that can shed light on ways to move the debate forward. If I am correct about these observations, then we can expect a fruitful theoretical discussion and an interesting reinterpretation of empirical evidence; if not, then the arguments against it will also bring to the surface a number of commitments that are at play in the debate.

Acknowledgements I would like to thank Dimitri Coelho-Mollo, Diana Pérez, Matteo Colombo, Guido Löhr and the anonymous reviewers who read this manuscript for their helpful comments and suggestions. I also thank participants of the "Examining scientific theories of emotion" symposium at the European Philosophical Society for the Study of the Emotions (EPSSE) Conference in Pisa, Italy (2019) and especially the co-organizers of the symposium Marco Viola and Rodrigo Díaz for their valuable feedback in previous presentations of this material. This article is part of my PhD project "Emotions as functional kinds: a meta-theoretical framework for the scientific study of the emotions" (Berlin School of Mind and Brain, Humboldt-Universität zu Berlin). I thank Jesse Prinz, Isabel Dziobek, and Michael Pauen for their supervision of this project.

Funding Information Open Access funding provided by Projekt DEAL.

References

Anderson, M.L. 2015. Mining the brain for a new taxonomy of the mind. *Philosophy Compass* 10 (1): 68–77.

Barrett, L.F. 2006. Are emotions natural kinds? *Perspectives on Psychological Science : A Journal of the Association for Psychological Science* 1 (1): 28–58.

Barrett, L.F. 2017. *How emotions are made: The secret life of the brain*. Boston: Houghton Mifflin Harcourt.

Brown, A.S., and E.J. Marsh. 2008. Evoking false beliefs about autobiographical experience. *Psychonomic Bulletin & Review* 15 (1): 186–190.

Celeghin, A., M. Diano, A. Bagnis, M. Viola, and M. Tamietto. 2017. Basic emotions in human neuroscience: Neuroimaging and beyond. *Frontiers in Psychology* 8.

Chalmers, D.J. 1997. *The conscious mind: In search of a fundamental theory*. New York: Oxford Univ. Press.

Colombetti, G. 2009. From affect programs to dynamical discrete emotions. *Philosophical Psychology* 22 (4): 407–425.

Colombetti, G. (2017). *The feeling body: affective science meets the enactive mind* (First MIT Press paperback edition). Cambridge, Massachusetts London, England: The MIT Press.

Crivelli, C., and A.J. Fridlund. 2018. Facial displays are tools for social influence. *Trends in Cognitive Sciences* 22 (5): 388–399.

Davitz, J.R. 1969. *The language of emotion*. New York: Academic Press.

Edelson, M., T. Sharot, R.J. Dolan, and Y. Dudai. 2011. Following the crowd: Brain substrates of long-term memory conformity. *Science* 333 (6038): 108–111.

Ekman, P. 1972. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation* 19: 207–282.

Ekman, P. 1980. Biological and cultural contributions to body and facial movement in the expression of emotions. In *Explaining emotions*, ed. A. Rorty, 73–102. Berkeley: University of California Press.

Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion* 6 (3): 169–200.

Ekman, P., and D. Cordaro. 2011. What is meant by calling emotions basic. *Emotion Review* 3 (4): 364–370.

Ekman, P., and W.V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17 (2): 124–129.

Ekman, P., E.R. Sorenson, and W.V. Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science* 164 (3875): 86–88.

Ekman, P., R.W. Levenson, and W.V. Friesen. 1983. Autonomic nervous system activity distinguishes among emotions. *Science* 221 (4616): 1208–1210.

Ekman, P., W.V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology* 53 (4): 712–717.

Elfenbein, H.A., and N. Ambady. 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* 128 (2): 203–235.

Frijda, N.H. 2007. *The laws of emotion*. Mahwah: Erlbaum.

Frijda, N.H., P. Kuipers, and E. ter Schure. 1989. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology* 57 (2): 212–228.

Gendron, M., D. Roberson, J.M. van der Vyver, and L.F. Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion* 14 (2): 251–262.

Griffiths, P.E. 1997. *What emotions really are: The problem of psychological categories*. Chicago: University of Chicago Press.

Hutto, D.D. 2012. Truly enactive emotion. *Emotion Review* 4 (2): 176–181.

Izard, C.E. 2007. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science* 2 (3): 260–280.

Izard, C.E. 2009. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual Review of Psychology* 60 (1): 1–25.

Jack, R.E., O.G.B. Garrod, H. Yu, R. Caldara, and P.G. Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109 (19): 7241–7244.

Jack, R.E., W. Sun, I. Delis, O.G.B. Garrod, and P.G. Schyns. 2016. Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General* 145 (6): 708–730.

Kassam, K.S., A.R. Markey, V.L. Cherkassky, G. Loewenstein, and M.A. Just. 2013. Identifying emotions on the basis of neural activation. *PLoS One* 8 (6): e66032.

Kragel, P.A., and K.S. LaBar. 2015. Multivariate neural biomarkers of emotional states are categorically distinct. *Social Cognitive and Affective Neuroscience* 10 (11): 1437–1448.

Kreibig, S.D. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84 (3): 394–421.

Lambie, J.A., and A.J. Marcel. 2002. Consciousness and the varieties of emotion experience: A theoretical framework. *Psychological Review* 109 (2): 219–259.

Lazarus, R.S. 1991. *Emotion and adaptation*. New York: Oxford University Press.

LeDoux, J.E. 2003. The emotional brain, fear, and the amygdala. *Cellular and Molecular Neurobiology* 23 (4–5): 727–738.

LeDoux, J.E. 2007. The amygdala. *Current Biology* 17 (20): 868–874.

LeDoux, J.E. 2012. Rethinking the emotional brain. *Neuron* 73 (4): 653–676.

LeDoux, J.E. 2013. The slippery slope of fear. *Trends in Cognitive Sciences* 17 (4): 155–156.

LeDoux, J.E., and R. Brown. 2017. A higher-order theory of emotional consciousness. *Proceedings of the National Academy of Sciences* 114 (10): E2016–E2025.

Levenson, R.W., P. Ekman, and W.V. Friesen. 1990. Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology* 27 (4): 363–384.
Levine, J. 1983. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64 (4): 354–361.
Lindquist, K.A., T.D. Wager, H. Kober, E. Bliss-Moreau, and L.F. Barrett. 2012. The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences* 35 (03): 121–143.
Loftus, E.F. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory* 12 (4): 361–366.
Mazzoni, G., and A. Memon. 2003. Imagination can create false autobiographical memories. *Psychological Science* 14 (2): 186–188.
Murphy, F.C., I. Nimmo-Smith, and A.D. Lawrence. 2003. Functional neuroanatomy of emotions: A meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience* 3 (3): 207–233.
Panksepp, J. 1998. *Affective neuroscience: The foundations of human and animal emotions*. Oxford: Oxford Univ. Press.
Panksepp, J. 2008. Carving "natural" emotions: "Kindly" from bottom-up but not top-down. *Journal of Theoretical and Philosophical Psychology* 28 (2): 395–422.
Panksepp, J. 2011. The basic emotional circuits of mammalian brains: Do animals have affective lives? *Neuroscience and Biobehavioral Reviews* 35 (9): 1791–1804.
Panksepp, J., and D. Watt. 2011. What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion Review* 3 (4): 387–396.
Pauen, M. 2017. The functional mapping hypothesis. *Topoi* 36 (1): 107–118.
Phan, K.L., T. Wager, S.F. Taylor, and I. Liberzon. 2002. Functional Neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* 16 (2): 331–348.
Phelps, E.A., and J.E. LeDoux. 2005. Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 48 (2): 175–187.
Price, C.J., and K.J. Friston. 2005. Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology* 22 (3–4): 262–275. https://doi.org/10.1080/02643290442000095.
Prinz, J.J. 2004. *Gut reactions. A perceptual theory of emotion*. Oxford: Oxford University Press.
Prinz, J.J. 2007. *The emotional construction of morals*. Oxford: Oxford University Press.
Rainville, P., A. Bechara, N. Naqvi, and A.R. Damasio. 2006. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology* 61 (1): 5–18.
Roseman, I.J. 2011. Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review* 3 (4): 434–443.
Roseman, I.J., C. Wiest, and T.S. Swartz. 1994. Phenomenology, behaviors, and goals differentiate discrete emotions. *Journal of Personality and Social Psychology* 67 (2): 206–221.
Russell, J.A. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin* 115 (1): 102–141.
Saarimäki, H., L.F. Ejtehadian, E. Glerean, I.P. Jääskeläinen, P. Vuilleumier, M. Sams, and L. Nummenmaa. 2018. Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience* 13 (5): 471–482.
Scarantino, A. 2012a. Functional specialization does not require a one-to-one mapping between brain regions and emotions. *Behavioral and Brain Sciences* 35 (03): 161–162.
Scarantino, A. 2012b. How to define emotions scientifically. *Emotion Review* 4 (4): 358–368.
Scarantino, A. 2015. Basic emotions, psychological construction, and the problem of variability. In *The psychological construction of emotion*, ed. L.F. Barrett and J.A. Russell, 334–376. New York: The Guilford Press.
Scarantino, A., and P.E. Griffiths. 2011. Don't give up on basic emotions. *Emotion Review* 3 (4): 444–454.
Siegel, E.H., M.K. Sands, W. Van den Noortgate, P. Condon, Y. Chang, J. Dy, et al. 2018. Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin* 144 (4): 343–393.
Smedslund, J. 1992. Are Frijda's "Laws of emotion" empirical? *Cognition & Emotion* 6 (6): 435–456.
Touroutoglou, A., K.A. Lindquist, B.C. Dickerson, and L.F. Barrett. 2014. Intrinsic connectivity in the human brain does not reveal networks for "basic" emotions. *Social Cognitive and Affective Neuroscience* 10 (9): 1257–1265.
Wittgenstein, L. (1953/2009). *Philosophical investigations* (Rev. 4th ed; P. M. S. Hacker & J. Schulte, Eds.; G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans.). Malden, MA: Wiley-Blackwell.
Yilmaz, M., and M. Meister. 2013. Rapid innate defensive responses of mice to looming visual stimuli. *Current Biology* 23 (20): 2011–2015.