COMMENTARY

emotion review

Emotion Review
Vol. 0, No. 0 (March 2022) 1–4

ISSN:1754-0739
DOI: 10.1177/17540739221089693
https://journals.sagepub.com/home/emr

Why a Connectionist Perspective on Emotion is Helpful

Gaurav Suri
Department of Psychology, San Francisco State University, San Francisco, California, USA

James J. Gross
Department of Psychology, Stanford University, California, USA

Abstract

To make progress related to long-standing questions related to the nature of emotion, we offer the Interactive Activation and Competition framework for Emotion (IAC-E). The IAC-E is not another conventional theory of emotion. Rather, it offers a neural-network-based, algorithmic account of how emotion instances and categories arise. Our approach suggests that there need not be a contradiction between instances of the same emotion being sometimes consistent and sometimes variable. Similarly, there need not be a contradiction between observations of homogeneity (common in the basic emotion approach) and heterogeneity (common in the constructed emotion approach) within emotion categories

Keywords
emotion, emotion categories, neural networks, connectionism, parallel distributed processing

In the target article (Suri & Gross, this issue), we argued that new insights regarding emotion are provided by the connectionist perspective, according to which all knowledge is resident in connections between simple processing units. We were therefore delighted that the distinguished scholars who provided commentaries on the target article saw promise in the connectionist approach and provided thoughtful insights which have led us to seek to more fully describe aspects of our Interactive Activation and Competition framework for Emotion (IAC-E).

We use this response to (a) situate the IAC-E within Marr's (1982) algorithmic level of analysis and describe the benefits of doing so, (b) elaborate on how the IAC-E can offer important perspectives related to the nature of emotion categories, (c) clarify the crucial role of the hidden units in the IAC-E, and (d) point to future algorithmic development of the IAC-E.

Benefits of Using the Algorithmic Level of Analysis in Emotion Research

The IAC-E framework is founded on the proposition that features of an emotion instance are bi-directionally connected to each other via conjunction (hidden) units. An instance of emotion occurs when input from one or more feature units produces activation in other units that it is connected to. The emergence of the emotion depends on the connections of its underlying neural network, and these connections may be innate and/or learned.

Crucially, the IAC-E does not offer yet another conventional theory of emotion. Rather, it offers a framework for thinking about *how* emotions and emotion categories might emerge. The IAC-E framework is thus situated within Marr's algorithmic level of analysis, which is concerned with describing the process-level mechanics of a phenomenon. As such, the IAC-E has a different target of explanation than other theories of emotion – several of which have been described in Moors' commentary (this issue) – because these theories of emotion are concerned with analyzing emotions at Marr's computational level of analysis.

Theories pitched at the computational level – which focuses on relations among high-level constructs – can be helpful in specifying the phenomena that require explanation. In the case of emotion, such theories are often predicated

Corresponding author: Gaurav Suri, Department of Psychology, San Francisco State University, San Francisco, California, USA.
Email: suri@sfsu.edu

upon assumptions related to the underlying functions of emotions, and the context and history within which they occur. These assumptions are in turn used to derive inferences about the fundamental nature of emotions and emotion categories. Unfortunately, differences among emotion theories are rampant, and these differences can often be traced to differences in underlying assumptions at the computational level of analysis. We believe that such differences frequently cannot be reconciled at the computational level alone. Such attempts often lead to each camp pointing to empirical phenomena that are consistent with their preferred theory, and ignoring phenomena that are not consistent with their preferred theory.

Theories formulated at the computational level often assume that their proposals can be implemented unchanged at lower levels. Such theories assume a "triumphant cascade" (Dennett, 1987, p. 227) through Marr's three levels so that specifications identified at the computational level (via considerations of the functions of emotion) are the main constraints at the algorithmic level and implementational level. We propose that there is no reason that descriptive specifications at the computational level should be favoured over constraints from other levels, and we believe that an exclusive focus on the computational level can lead to theoretical accounts that are incomplete and difficult to compare with other accounts.

In the IAC-E, we place the greatest emphasis at the process (algorithmic) level, while allowing that the representations and processes at this level are interdependent with computational level constraints related to the functions of emotions and the context and history within which they occur. Like other connectionist frameworks, the IAC-E is also sensitive to implementation level constraints imposed by our biology. With its openness to a broader set of constraints, we see our emergentist framework as providing new insights to core questions about the nature of emotion.

For example, emotion theories developed at the computational level sometimes differ on whether instances of the same emotion are generally consistent or generally variable. This has framed the debate in a manner that suggests that only one of these options must be true. The algorithmic processes of the IAC-E suggest that emotions are – under some model parameter values – generally consistent, and under other model parameter values, quite variable. From the algorithmic perspective of the IAC-E, these two observations need not be in conflict.

Importantly, the IAC-E is agnostic as to whether a particular set of features should or should not be included in the definition of an emotion. We believe that such issues depend on researcher objectives and cannot be "right" or "wrong" in a theory-independent sense. However, the IAC-E does provide explicit process-level details about how any emotion variable – including a goal-related variable (see Moors, this issue) – might interact with other emotion-related variables that are considered important given the investigator's objectives, and what the consequences of such interactions are likely to be.

Elaborating the IAC-E Perspective on the Nature of Emotion Categories

Another benefit of the algorithmic approach of the IAC-E is that it can reconcile differing intuitions about the relative homogeneity or heterogeneity of emotion categories.

Proponents of basic emotion views – either implicitly or explicitly – subscribe to the classical view of categorization (described in Medin, 1989) which assumes that the mental representation of a category consists of a summary list of features that are each necessary for category membership and are jointly sufficient for ensuring category membership. For example, the category of prime numbers consists of numbers that are greater than 1, and whose only divisors are 1 and the number itself. To see if a number belongs to the category of prime numbers, one must ensure that both these properties are present. Conversely, if either of these properties is not present, then the number is not in the category of primes. According to basic emotion views, an emotion instance belongs to an emotion category if, and only if, its unfolding is accompanied by the firing of a neural circuit that is specific to that family of emotion instances. For example, an emotion instance may be classified as fear if, and only if, the neural circuit corresponding to the fear category is activated.

Proponents of constructed views of emotion – either explicitly or implicitly – subscribe to the knowledge-based view of categorization (Murphy & Medin, 1985) which asserts that categories emerge when people use their knowledge of the underlying domain to categorize new instances. For example, Barsalou (1983) argued that children, pets, photo albums, family heirlooms, and cash would not ordinarily appear to belong to the same category. However, when one imposes knowledge related to things that are precious and must be rescued when the house is on fire, this collection comprises a meaningful category into which new instances can be assigned (or not assigned). According to the constructed view, imposed knowledge is crucial for emotion categories to emerge.

We observe that the exemplar method of categorization – a widely used alternative to the classical and knowledge-based categorization (Medin, 1989) – can be productively applied within the IAC-E to explain category emergence. According to the exemplar-based categorization method, a new emotion instance is categorized with the prior emotion instance that it most resembles. Since the exemplar method depends on the grouping of like instances, its application in the emotion domain requires a criterion to determine whether or not instances of emotion – i.e., conjunctive units in the hidden pool – are similar to each other. One promising way to accomplish this is to use the similarity of component features that are connected to each of the conjunctive units. To build an intuition for this process, we use an illustrative example.

Say Hidden Unit 1 (HU1) is connected to feature (input) units representing heart rate and approach/avoid behaviour. Let us assume that the component units for HU1 are a heart rate of 100 and rapid approach behaviour. Similarly assume that the component units for HU2 are a heart rate of 60, and slow avoid behaviour and the component units of HU3 are a heart rate of 88 and rapid approach behaviour. Intuitively, we would want to claim that HU1 and HU3 are similar to one another and dissimilar from HU2. This intuition can be extended to compute similarity scores between any two hidden units – even when they are connected to feature units in different pools.

If an experimenter includes a small number of feature categories in her observations of emotion, the similarity scores within each category will, in general, be greater than if she had included a greater number of feature categories. Similarly, if an experimenter limits her scope to features (within a feature category) that tend to have tight covariances, then the similarity scores within each category will tend to be greater than if she had included a broader set of features.

The IAC-E approach suggests that there need not be a contradiction between observations of homogeneity and heterogeneity within emotion categories. These categories are emergent consequences of the consideration sets that are used to generate the categories in the first place. The algorithmic processes of the IAC-E, in conjunction with the exemplar approach of categorization, can demonstrate that different assumptions about what is, and is not, included in consideration sets will lead to the emergence of different types of categories.

The Role of the Hidden Units in the IAC-E

Comments from Lench and Reed (this issue) have encouraged us to clarify the role of the hidden units in the IAC-E. All hidden units are simply conjunction units between feature units. As such, they do not represent any single feature, and they can be activated by activations in multiple feature units. If we label feature units (as we do in some simulations in the target article), we do so for the sake of clarity in describing the particular network connections. The purpose of hidden units is to enable consistency of emotion response patterns, malleability of emotion responses over time, and context sensitivity of emotion responses in different eliciting scenarios. We next describe each of these attributes enabled by the hidden units.

The interactivity enabled by hidden units can lead to *consistency* of emotion response patterns via at least two mechanisms. First, consistency may occur due to the presence of hidden units that innately connect selected feature input units to each other. In such a situation, if a feature pool unit (say unit A) receives external input, it activates a population of hidden units it is innately connected with; these units in turn activate the other feature units they are innately connected with (say units B and C). If there are no context-related feature units to divert activation, then activation in unit A will reliably produce activation in units B and C.

The second mechanism enabling consistency relies upon statistical regularities in the environment to produce a population of well-connected hidden units. Here, hidden units do not innately connect feature units. Instead, hidden units are developed over time in consistently co-occurring populations of input units. Such units can produce similar consistency as innate units.

Connecting feature units to each other via conjunctive hidden units affords the opportunity to explain *malleability* of emotion-related responses over time. For example, consider the case of a person who has experienced many instances of emotions in which a higher-than-normal level of somatic arousal is associated with withdraw motivation (corresponding, for example, to fear) and only a few instances of emotions in which a higher-than-normal level of somatic arousal is associated with approach motivation (corresponding, for example, to excitement). If such a person experiences a high heart rate (an activation of units in the somatic input pool), she is likely to exhibit increased activation in withdraw motivation units. In the IAC-E, this occurs because the 'increased heart rate' input unit activates many prior hidden units – each of which send activation to the 'withdraw' motivation unit. Contrastingly, fewer hidden units send activation to the 'approach' motivation unit.

However, this association is malleable: imagine that this person next encounters several other instances of emotion that are associated with a high heart rate, but are also associated with an approach motivation (e.g., because she frequently rides on an enjoyable Ferris wheel). These emotion instances form new conjunction units in the hidden pool, each representing an instance of excitement. Now the 'high heart rate' feature units would activate conjunction units that send activation to 'approach' (as well as earlier conjunctions sending activation to 'withdraw'). This would effectively weaken the association between experiencing a high heart rate and wanting to withdraw. With enough new hidden unit conjunctions, a high heart rate may even become associated with approach motivation. Hidden units thus play the role of casting top-down votes that determine the effective associative strength between different emotion features.

Finally, associations involving hidden units are subject to selective activation based on combinations of input cues (McClelland, 1981; Medin & Schaffer, 1978) which in turn causes *context sensitivity* in the IAC-E. Increasing specificity of input can alter the pattern of activated conjunctions in the hidden pool. Continuing with our example above, providing input into the 'high heart rate' and 'withdraw' feature units will activate hidden units that often connect feature units corresponding to instances of fear and disgust; this activation may in turn activate the feature units associated with the subjective experience of fear or disgust. On the other hand, input into the 'high heart rate' and 'approach' feature units will activate the hidden unit that connects feature units corresponding to instances of excitement; this activation may in

turn, activate the feature units associated with the subjective experience of excitement.

In each of these examples, consistency of response, malleability, and context sensitivity are emergent consequences of particular starting conditions and subsequent inputs. An instance of emotion is precisely the spread of activation in the units of the IAC-E. This spread is a transparent consequence of the starting inputs and the connection weights of the network.

Moving Beyond Localist Networks

The IAC-E is a localist network in that it represents a single cognizable emotion feature in a single unit. Localist networks may be thought of as approximate characterizations of more complex distributed networks. They provide the convenience of being able to render all of the dynamics in terms of conceptual entities, rather than in terms of the individual neuronal-level dynamics. They are therefore easier to comprehend than equivalent distributed networks. In a sense there is continuum of networks at the algorithmic level of analysis with localist versions being closer to the computational level and distributed versions being closer to the implementation level.

Individual connections between units in a localist network stand in for interpretable associative relationships. For example, in the IAC-E, individual (indirect) connections between feature units implies that activation in one feature will reliably produce activation in the connected feature. In distributed models, however, the situation is more complex. In such systems, if one wishes to associate, for example, the 'elevated blood-pressure' with the 'avoid' action tendency, and if these features are each represented as a pattern of activation over a set of units, then the connection weight changes required to store the association may involve many, or even all, of the weights involved in other associations (McClelland & Cleeremans, 2009).

Although the localist IAC-E network presented in our target article captures many of the essential algorithmic dynamics we sought to understand, we believe that there is a promising opportunity to create versions of the IAC-E that instantiate distributed representations (Plaut & McClelland, 2010). Distributed networks offer several advantages over localist representations: first, they tend to preserve latent regularities in the input such that items with similar properties tend to have similar representations. In the context of emotion, this property is particularly useful because it would allow one to estimate the degree of overlap in input properties of different emotion instances (within and across different categories) via the extent of overlap in their neural (distributed) representations. Such calculations are not possible in localist networks. Second, distributed representations offer opportunities to simulate the gradual degradation of emotional responding in diseases such as frontotemporal dementia (Marshall et al., 2019). Distributed networks may be lesioned (by adding noise to activation transfer), and the effects of such lesioning may be compared to empirical results related to patient populations. Finally, and importantly, a distributed version of the IAC-E offers enhanced opportunities for constraints from Marr's (1982) implementation level of analysis. We believe that including constraints from multiple levels analysis will prove to be generative in emotion research, and in psychology in general.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Gaurav Suri https://orcid.org/0000-0002-0423-060X

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, *11,* 211–227. https://doi.org/10.3758/BF03196968

Dennett, D. (1987). *The intentional stance*. The MIT Press.

Lench, H. C., & Reed, N. T. (in press). Can we model what an emotion is? Comment on Suri & Gross. *Emotion Review*.

Marr, D. (1982). *Vision*. Freeman.

Marshall, C. R., Hardy, C. J., Russell, L. L., Bond, R. L., Sivasathiaseelan, H., Greaves, C., Moore, K. M., Agustus, J. L., van Leeuwen, J. P., Wastling, S. J., Rohrer, J. D., Kilner, J. M., & Warren, J. D. (2019). The functional neuroanatomy of emotion processing in frontotemporal dementias. *Brain*, *142,* 2873–2887. https://doi.org/10.1093/brain/awz204

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Conference of the Cognitive Science Society* (pp. 170–172).

McClelland, J. L., & Cleeremans, A. (2009). Consciousness and connectionist models. In T. Bayne, A. Cleeremans, & P. Wilken (Eds.), *The Oxford companion to consciousness* (pp. 180–181). Oxford University Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44,* 1469–1481. https://doi.org/10.1037/0003-066X.44.12.1469

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85,* 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Moors, A. (in press). Commentary: Old wine in new bags—Suri and Gross's connectionist theory of emotion is another type of network theory. *Emotion Review*.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92,* 289–316. https://doi.org/10.1037/0033-295X.92.3.289

Plaut, D. C., & McClelland, J. L. (2010). Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review*, *117,* 284–288. https://doi.org/10.1037/a0017101

Suri, G., & Gross, J. J. (in press). What is an emotion? A connectionist perspective. *Emotion Review*.