WHEN AND WHY TWO EMOTIONAL EXPERIENCES ARE SIMILAR TO EACH OTHER

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2022

Martina Riberto

Faculty of Biology, Medicine and Health,
Division of Neuroscience and Experimental Psychology

Table of Contents

Number of words: 52221

Table of Figures

Table of tables

Abstract

Emotional similarity refers to the tendency to group stimuli together according to the feelings they evoke in us. The study of emotional similarity is relevant for the semantic memory research and overgeneralisation bias in anxiety disorders and may have impact on psychological well-being. Most of the studies on similarity have focused on non-emotional stimuli, and fewer on simple stimuli (e.g., shapes, objects) that acquired an emotional value following fear conditioning. Very little is known about what makes us perceive real-life emotional experiences as similar. We assumed a similarity space of several integrated dimensions, with emotions as the most influential. We predicted emotional stimuli to be judged as more similar to each other than neutral stimuli, as they shared low scores in valence and high scores in arousal, and thus they were more salient. We also expected this to be underpinned by higher similarity in the neural activation patterns of emotional than neutral stimuli. We combined different similarity judgements tasks and fMRI to test our hypotheses, and analysed the data using RSA. Our results suggest an important role of emotion on similarity perception. Even though two expressions of the same person were objectively more similar to each other than the faces of two different individuals who expressed the same emotion, participants judged both types of face pairs to be just as similar to each other. However, the similarity between faces expressing similar emotions was lower than between neutral faces. In addition, we found no differences between images of the two emotional than the two neutral categories of real-life events. Similar findings were replicated using stimuli that acquired an emotional value after aversive conditioning, suggesting that emotion is as relevant as visual and semantic dimensions in perceived similarity. Despite this equivalence in similarity perception, emotions were more influential in the neural similarity space, resulting in higher similarity among the neural representations of emotional compared to neutral stimuli. We observed this in brain clusters located in the ventral visual stream underlying semantic processing and categorisation, and in regions involved in affect representations (i.e., precuneus, insula) and modulation (i.e., dorsal anterior cingulate cortex, dorsomedial prefrontal cortex). This pattern of findings suggest that emotions might trigger local (within a brain region) and distant (between brain regions) synchronisation processes, such that a stable mental representation, which

encoded the ‘relevance’ of the stimulus, emerges and is shared among emotional experiences.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made**.**

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy, in any relevant Thesis restriction declarations deposited in the University Library, the University Library’s regulations and in the University’s policy on Presentation of Theses.

For Maria Cristina

Acknowledgements

To my supervisors who believed in me since the beginning, unpacked my thoughts patiently, guided me to fulfil my potential during these years with support, sensitivity, trust, and who taught me that science is not everything in life.

To my donor and the Weizmann UK that gave me this unique life opportunity and without whom nothing would have been possible.

To my lab mates who warmly welcomed me in the early fragile moments, who laughed with me about my “Italianish”, who advised and supported me, and patiently assisted me during very frequent Matlab crises.

To the most brilliant scientist I know, who initiated my passion, taught me everything I know in Neuroscience, and inspired my creativity during my PhD.

To my family who accompanied me during this long journey with words of love and encouragement, always reminding me that there is only one home.

To my old friends who kept on being my friends although the pain of seeing me leaving every time, who patiently dealt with my crazy visits schedule, craving for the next return, and who made me feel loved despite the distance between us.

To all of those who became dear friends during this journey with whom I have always felt at home and who made me understand that family is what you choose to be.

To all the people I met on this path thanks to whom I appreciated the beauty of human diversity.

To all the participants that contributed to my research, who found it fascinating, making me feel that what I was doing was important.

To Tale of Us who eased many of my days with their melodies.

To the one who enlightened my way and showed me the best part of me.

To Maria Cristina who keeps on being “one of my cornerstones” and, with this, encouraged me to resist when I wanted to give up.

I am grateful.

Rationale

Emotional similarity refers to the tendency to group stimuli together because they evoke the same feeling in us. Emotional similarity is a fundamental principle in cognition, as it supports core cognitive functions, such as categorisation, semantic memory and learning. Research on emotional similarity may have clinical implications in anxiety disorders. After a traumatic event, patients overgeneralise, as they consider later experiences as similar to the original fearful one not because of their ostensible meaning, but their emotional similarity. In addition, it might help to shed more light on the neurobiological mechanisms underlying semantic cognition.

In this thesis, we explored when and why we judge two emotional experience to be similar to each other, and whether we represent them differently from non-emotional experiences. We assumed a similarity space that comprises of several integrated visual, semantic and emotional dimensions, with the latter being the most influential to overall perceived similarity. We also expect this to be underpinned by higher similarity among neural representations of emotional than neutral stimuli. This would enable individuals to differentiate emotionally relevant stimuli from those that are not as function of survival.

In **chapter 1**, we summarised major findings in similarity judgements research, focusing on the semantic (*section 1.2*) and emotional (*section 1.3*) facets of similarity. We also highlighted the main limitations in the similarity judgements literature, and proposed different ways to overcome them in future studies (*section 1.4).* This literature review has been published in *Brain Topography* (2019). In **chapter 2**, we revised the experimental procedures and data analyses mostly used in similarity literature. We first described the stimulus sets generally used in similarity tasks, by focusing on visual stimuli (*section 2.1*) and the experimental procedures, whereby direct and indirect similarity estimates can be derived (*sections 2.2 - 2.3*). Finally, we defined neural similarity in the context of Representational Similarity Analysis (*section 2.4*), concluding with a summary of the statistical analyses we conducted on behavioural and neural data (*section 2.5*). In **chapters 3**, **4**, **5** we tested our hypotheses using two similarity judgements tasks (i.e., pairwise ratings and multi-arrangements tasks) and several databases of

stimuli that differed in visual complexity (e.g., pictures of animals, complex scenes) and on emotional dimensions (e.g., fearful faces; scenes of car accidents). Particularly, in **chapter 3** we studied whether emotionally similar faces were rated as more similar than neutral faces. We published the results of this experiment in the journal *Symmetry* (2021). The same hypothesis was tested using real-world photographs that differed in valence and arousal in **chapter 4**. Specifically, in experiment 1, we tested our main hypothesis using two datasets of stimuli, the second of which allowed us an exquisite control over semantic similarity. In experiment 3, we also investigated whether the neural representations of emotional stimuli were more similar than those associated with neutral stimuli. This full chapter was adapted as paper and it has been accepted for publication in the Journal of Neuroscience (2022). Finally, in **chapter 5** we explored whether aversive conditioning was associated with an increase in behavioural (experiment 1) and neural (experiment 2) similarity measures between threatening than not threatening images and sounds of animals. In **chapter 6**, we discussed our results and their impact on memory and semantic cognition literature as well as in anxiety disorders.

1. Chapter: Introduction

This chapter has been published in *Brain Topography* (Martina Riberto, Pobric, & Talmi, 2019).

'A group of people, books, whether of a certain kind and certain states of mind are all grouped together as alike [...] What holds them together […] is the evocation of a defining affective response' (Bruner & Austin, 1986).

Emotional similarity refers to the tendency to group stimuli together according to the feelings they evoke in us. It is a fundamental principle in cognition, as it supports core functions, such as categorisation (Barrett, 2017; Barsalou, 2017), memory and learning (Leal, Tighe, Jones, & Yassa, 2014; Leal & Yassa, 2018; Talmi & McGarry, 2012). Research on emotional similarity may have clinical implications for the overgeneralisation bias in anxiety disorders. After a traumatic event, patients consider later experiences as similar to the original fearful one not because of their ostensible meaning, but their emotional similarity (Ahrens et al., 2016; Laufer, Israeli, & Paz, 2016). The majority of research on similarity perception that has been conducted to date has focused on non-emotional stimuli. Different models have been proposed to explain how we represent semantic concepts, and judge the similarity among them. However, very little is known about what makes us perceive real-life emotional experiences as similar.

In this chapter, we first introduced the construct of emotional similarity, the possible relationships with semantic similarity, its relevance in memory research and anxiety disorders (*section 1.1*). Second, we reviewed the literature about semantic (*section 1.2*) and emotional (*section 1.3*) similarity, focusing on neuroimaging studies aimed at developing neurobiological model of semantic cognition and emotional categorization. Third, we highlighted the limitations in emotional similarity literature, possibly due to confounding factors during stimulus selection process (*section 1.3.1*). Finally, we proposed future directions in emotional similarity research to improve our understanding of the cognitive and neural mechanisms of this core construct. A multi-modal and overarching approach, which combine behavioral and neural data, would be the key to further unveil what makes emotional experiences similar to each other (*section 1.4*). This

chapter has been published as literature review in Brain Topography (Martina Riberto et al., 2019) and served as general background for the experimental chapters (chapters 3, 4 and 5).

1.1 Background

Emotional similarity refers to the similarity between the feelings that stimuli evoke in us. Poets and storytellers routinely use the power of emotional similarity to convey the emotional tone of a situation by analogy, for example, when the sadness that follows the breakup of a relationship is likened to that we feel when the weather is bad. As the famous song goes, it is 'stormy weather, since my man and I ain't together, keeps raining all the time…'. According to Bruner, stimuli that are very different visually and semantically may nevertheless be perceived as similar to each other because of the feelings they evoke in us (Bruner & Austin, 1986). For example, we may judge an image of a homeless person begging for food and an image of a businesswoman talking on the phone as different, even if the pictures are taken at the same street corner, because one evokes a negative feeling and one a neutral feeling. On the contrary, the same image of a beggar and an image of a person injured in a car accident may be evaluated as more similar if both evoke negative feelings, even if the pictures are taken in different places around the world. In Bruner's discussion, emotional similarity is considered an orthogonal dimension to the visual and semantic dimensions of a stimulus. Alternatively, the emotional facet of our experience of a stimulus may be considered part of its semantic meaning; in that case, emotional similarity may be reduced to a specific form of semantic similarity.

We define emotional similarity as the similarity between the emotional dimensions of stimuli in the representational space. This space is in part objective and shared among individuals, and in part subjective and in continuous interaction with our experience.

The majority of research on similarity perception that has been conducted to date has focused on non-emotional stimuli, such as words, object, shapes, faces and scenes. In these studies (Goldstone, Medin, & Halberstadt, 1997; Golonka & Estes, 2009; Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2014; Iordan, Greene, Beck, & Fei-Fei, 2015; King, Groen, Steel, Kravitz, & Baker, 2019), participants

were involved in explicit similarity judgement tasks. In others (Bruffaerts et al., 2013; Clarke & Tyler, 2014; Guntupalli, Wheeler, & Gobbini, 2016; Haxby et al., 2001; Haxby et al., 2011; N. Kriegeskorte, M. Mur, & P. A. Bandettini, 2008b; Neyens et al., 2017), the main interest was the neural similarity, namely the similarity among neural representations associated with non-emotional stimuli during tasks not related to the similarity judgements. By contrast, less is known about what makes people perceive richer, life-like events as similar, and even less when these are emotional. Understanding the cognitive and neural mechanisms underlying emotional similarity may have implications for research on categorisation (Barrett, 2004, 2017; Barsalou, 2017), memory of emotional experiences (Leal, Tighe, Jones, et al., 2014; Leal & Yassa, 2018; Talmi & McGarry, 2012), and generalisation (Dunsmoor, Kragel, Martin, & LaBar, 2013; Laufer & Paz, 2012; Schechtman, Laufer, & Paz, 2010). From a clinical perspective, the study of emotional similarity could help in understanding why patients with anxiety disorders overgeneralise and judge a variety of subsequent experiences to be similar to the original fearful one (Laufer et al., 2016; Lissek et al., 2009).

Below, we review the major findings and debates in the literature on similarity, with the goal of placing the concept of 'emotional similarity' within the context of relevant research. With this aim, we will summarise two lines of research, one focused on explicit similarity judgements and the other on neural similarity. This is because both of them provide interesting information about what makes two stimuli similar, in terms of both cognitive dimensions and neural mechanisms. First, we will focus on semantic similarity, namely the similarity among non-emotional stimuli. We will use this literature as background for the emotional facet of the similarity, and to ask how the emotional similarity could be incorporated. Is emotional similarity a facet of semantic similarity or is a further dimension in a complex semantic space? We will end by proposing future directions in this research field.

1.2 Semantic similarity

We may judge two stimuli, such as a blue circle and a blue ellipse, as similar, because they share some features (the rounded shape and blue colour). Because

of the number of properties that they share, we will consider them more similar than a blue ellipse and a pink square. This is line with the 'contrast model', which posited that similarity between two items is a function of their common features weighed against their distinctive features (Tversky, 1977). The 'contrast model' is limited in that it fails to consider the relationships among features (Gentner & Markman, 1994, 1997; Markman & Gentner, 1993). These include thematic and taxonomic relationships, which widely contribute to semantic memory and similarity judgements (Hoffman, Ralph, & Rogers, 2013; E. L. Lin & Murphy, 2001; M. A. L. Ralph, Sage, Jones, & Mayberry, 2010; Schwartz et al., 2011).

Milk paired with *jam* is an example of thematic relationship. Thematic relationships are defined as any temporal, spatial, causal, or functional relationships between objects, which perform complementary roles in the same scenario or event (e.g., *breakfast*) (Estes, Golonka, & Jones, 2011). It is widely known in the semantic memory literature that people judge thematically related stimuli to be more similar to each other than other stimuli (Chen et al., 2013; Estes et al., 2011; Golonka & Estes, 2009; Simmons & Estes, 2008). The paradigmatic stimuli are natural, complex pictures (Lang, Bradley, & Cuthbert, 2008; Marchewka, Żurawski, Jednoróg, & Grabowska, 2014). For these stimuli, thematic relationships can arise from affordances (Maguire, Maguire, & Cater, 2010), namely the possible actions that a person can perform in a specific situation. As shown by Greene et al. (2014), affordances may even be the most salient dimension in the categorisation of natural scenes. In that study, participants categorised natural complex pictures mainly according to affordance, rather than visual or taxonomic similarity (Greene et al., 2014).

Labrador and *Chihuahua* are taxonomically similar. While visually these animals are different (different colour, size, etc.), they are also similar, because they share some features (both bark and are four-legged), which once related bring out the category *dogs.* Thus, we group these items in the same category, *dogs*, and judge them as more similar than items from different categories (Chen et al., 2013; Wisniewski & Bassok, 1999; Xiao, Dong, Chen, & Xue, 2016; Xu et al., 2018). People also generalise these properties to new items with similar features (e.g., *German Shepherd*), and attribute to these items extra features that define the category, even if those were never directly experienced (Jackson, Hoffman, Pobric, & Lambon Ralph, 2015). Features-based categories are organised

hierarchically in semantic memory (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Within this hierarchy, it is often possible to distinguish between different levels: the broadest level is the superordinate (e.g., *mammals*), then the basic (e.g., *dogs*) and then the subordinate (e.g., *Labrador*). Although some examples do not fit this neat classification (e.g., screwdriver or lawnmower) and there are a number of contradictory findings (Rogers & Patterson, 2007; Taylor, Devereux, Acres, Randall, & Tyler, 2012), many studies showed that participants are more accurate and faster in categorising objects at the basic level than at the superordinate and the subordinate level (Anglin, 1977; Horton & Markman, 1980; Iordan et al., 2015; Mack, Wong, Gauthier, Tanaka, & Palmeri, 2009; G. L. Murphy & Brownell, 1985). Many of the stimuli in the emotional cognition literature have taxonomic relationships. In the IAPS set, for example, a picture of a man pointing a gun and a man wielding a knife are subordinates of the basic level 'aggravated assault'. Emotional events are the core of our life stories and their categorisation, as well as the similarity among them, are fundamental to make them meaningful. However, most of the studies focused on the neural mechanisms underlying the perception of similarity among neutral stimuli and on the neural representations of non-emotional stimuli during cognitive and perceptual tasks.

1.2.1 Neuroimaging studies

It is possible to map in the brain the similarity structure observed at behavioural level, by using innovative Multivariate Pattern Analysis (MVPA) methods. Among them, Representational Similarity Analysis (RSA) gained popularity in neuroscience in the last decade to investigate the cognitive and neural mechanisms of perceived similarity. This technique allows combining neural evidence with behavioural and computational data by calculating their correlation. In this way, it is possible to test whether and where the similarity structure observed at behavioural level is represented in the brain. In addition, this correlational-based technique examines the correlation between the neural representations of stimuli, as it is measured through the BOLD signal during cognitive tasks in fMRI, to draw conclusions about their similarity (Kriegeskorte & Mur, 2012; Kriegeskorte, Mur, et al., 2008b; Nili et al., 2014). In a recent MVPA study, Iordan et al. (2015) explored how the different levels of semantic categories

are represented across the occipitotemporal cortex. They hypothesised that categorisation may be an emergent property of the human ventral visual system. In order to test this hypothesis, they calculated the category boundary effect as the difference between cohesiveness (within-category neural similarity) and distinctiveness (between-categories neural similarity). This quantity provides a measure of how well categories are separated at each taxonomic level. For example, at the basic level, cohesion for 'dogs' is defined as the average correlation between voxel activations associated with the presentation of a 'dog' and any other type of 'dog'. On the other side, at the basic level, distinctiveness for 'dogs' is defined as average correlation between voxel activations associated with the presentation of a 'dog' and, for example, a 'flower'. They found high cohesiveness in V1, such that patterns elicited by subordinates are not distinguishable. As we move up in the ventral visual stream (i.e., lateral occipital cortex, posterior middle temporal gyrus, inferior temporal gyrus), the categories become more sharply distinguishable at basic level (Iordan et al., 2015). This is in line with other studies, which showed that inferior temporal regions are involved in semantic categorisation and perceived similarity of objects (Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014; R. Epstein & Kanwisher, 1998; Grill-Spector, Kushnir, Edelman, Itzchak, & Malach, 1998; Kriegeskorte, Mur, Ruff, et al., 2008; Malach et al., 1995; Martin, Wiggs, Ungerleider, & Haxby, 1996) and faces (Guntupalli et al., 2016; Haxby et al., 2001; Haxby et al., 2011). Thus, according to these studies, semantic knowledge is not 'located in' one brain area, but it arises from distinct patterns of response that are distributed across brain regions (Haxby et al., 2001).

A similar perspective is reflected in the 'hub and spoke' model (Rogers et al., 2004) , an influential model of semantic memory. According to this model, semantic categorisation is the result of an interaction between different modality-specific cortices (i.e., the 'spokes') distributed across the brain, and a transmodal 'hub', located in the ventral part of the anterior temporal lobe (vATL) (Lambon Ralph, 2014; Patterson, Nestor, & Rogers, 2007; M. A. L. Ralph et al., 2010). In particular, the 'hub' integrates sensory, motor and verbal information that together define a concept, and which are encoded in the different 'spokes'. It also extracts inter-stimulus relationships that go beyond visual similarities, such as taxonomic and

thematic relationships, and generalise these relationships to new stimuli with similar features. Many neuropsychological and neuroimaging findings, both in patients with semantic dementia (Bozeat, Ralph, Patterson, Garrard, & Hodges, 2000; Guo et al., 2013; Jefferies, Patterson, Jones, Ralph, & Matthew, 2009; P. J. Nestor, Fryer, & Hodges, 2006; M. L. Ralph, Lowe, & Rogers, 2007) and in healthy controls (Pobric, Jefferies, & Ralph, 2007; M. Visser, Jefferies, Embleton, & Lambon Ralph, 2012) support this model. The vATL interacts also with other brain regions, which are part of the semantic control (SC) network, to generate context-dependent semantic representations. This network include the posterior middle temporal gyrus, the prefrontal cortex, the intraparietal sulcus, the pre-supplementary motor area and the anterior cingulate cortex (for a review on this topic, see (M. A. L. Ralph, Jefferies, Patterson, & Rogers, 2017)). Finally, as reviewed by Rice et al. (2018) the ATL is also involved in processing socially relevant semantic concepts, including person face knowledge and emotional concepts (Collins & Olson, 2014; Olson, McCoy, Klobusicky, & Ross, 2013; Pobric, Lambon Ralph, & Zahn, 2016; Wang et al., 2017; Zahn et al., 2009; Zahn et al., 2007), because of its connection with the amygdala and orbitofrontal regions through the uncinate fasciculus (Highley, Walker, Esiri, Crow, & Harrison, 2002; Von Der Heide, Skipper, Klobusicky, & Olson, 2013). These regions might be thought as 'emotional spokes', which interact with the ATL to generate emotional concepts. Future studies are needed to test this hypothesis.

To summarise, semantic similarity supports core cognitive functions, such as semantic categorisation and semantic memory. Recent neuroimaging findings showed that conceptual knowledge is a widely distributed neural network, which include occipitotemporal and prefrontal regions. Different models have been proposed to explain the cognitive and neural mechanisms of semantic knowledge and similarity judgments (Caramazza, Hillis, Rapp, & Romani, 1990; A. R. Damasio, 1989; Riddoch, Humphreys, Coltheart, & Funnell, 1988). However, to our knowledge, these perspectives are limited to non-emotional stimuli, and have never been tested in the context of emotional similarity and categorisation.

1.3 Emotional similarity

While the majority of the studies about similarity judgements focused on non-emotional stimuli, a vast literature in emotion research asks what makes two emotional stimuli similar. To answer this question, participants are often asked to sort simple stimuli, such as words or faces, according to their similarity, or to rate the similarity among them on a Likert scale (Calvo & Nummenmaa, 2008; Jamin Brett Halberstadt & Niedenthal, 1997; Jamin B Halberstadt, Niedenthal, & Kushner, 1995; Koch, Alves, Krüger, & Unkelbach, 2016; Osgood, 1952; Roberts & Wedell, 1994; Russell & Bullock, 1985; Russell & Pratt, 1980; Said, Moore, Engell, Todorov, & Haxby, 2010; Schlosberg, 1952; van Tilburg & Igou, 2017). The paradigmatic finding is that participants judge the similarity according to two dimensions, the valence and the arousal of the stimuli. These dimensions are not explicitly used during the similarity judgements, but rather they represent implicit components of the cognitive structure underlying these stimuli (Barrett, 2004). We can map this cognitive structure by using Multidimensional Scaling (MDS) procedure. When represented in a geometric space, defined by valence and arousal as orthogonal axes, emotional stimuli are placed along the perimeter of a circle. This is the core idea of Russell's 'circumplex model of emotion' (Russell & Pratt, 1980) and other dimensional theories of emotion (Bradley, Greenwald, Petry, & Lang, 1992; Mehrabian, 1980; R. Plutchik, 2001; Watson & Tellegen, 1985), which have been widely used in emotion research (Barrett & Russell, 1999; A. Damasio, 2003; Kuppens, Tuerlinckx, Russell, & Barrett, 2013; Lang et al., 2008; Mäntylä, Adams, Destefanis, Graziotin, & Ortu, 2016; Marchewka et al., 2014; Yu et al., 2016; Zevon & Tellegen, 1982). In this representational space, the distance among stimuli reflects their similarity, with short distances representing high similarity. The multi-arrangement method, a direct way to measure similarity, is based on this principle (Kriegeskorte & Mur, 2012). This quick and efficient task is used for experiments with a relatively large set of stimuli, because participants simultaneously judge the similarity among many stimuli displayed together (Chavez & Heatherton, 2015; Chikazoe, Lee, Kriegeskorte, & Anderson, 2014), as opposed to a pairwise presentation.

Emotional similarity can be also quantified indirectly. Asking participants to rate the semantic relatedness between words (Talmi & Moscovitch, 2004) or pictures

(Gallo, Foster, & Johnson, 2009; Sison & Mather, 2007; Talmi, Luk, McGarry, & Moscovitch, 2007; Talmi & McGarry, 2012) is an example of an indirect measure of similarity. This is because the higher the relatedness between concepts in semantic memory, the higher the similarity between them. These studies suggest that emotions increase the semantic relatedness, resulting in higher ratings among negative emotional stimuli compared to neutral ones. This might lead to a better organisation of emotional stimuli, and might explain the advantage they have in immediate memory tests (Talmi, 2013; Talmi & McGarry, 2012).

The findings above indicate that emotion increases perceived similarity between stimuli. Greater perceived similarity among emotional stimuli might be related to the effect of arousal on hippocampal pattern separation, the ability to store similar experiences in distinct and non-overlapping representations. This might explain why participants find it harder to discriminate between targets and similar lures when those are emotional (Leal, Tighe, Jones, et al., 2014; Leal & Yassa, 2018; Mattar & Talmi, 2019; Segal, Stark, Kattan, Stark, & Yassa, 2012; Zheng et al., 2019). Other studies suggested that the arousal might also increase the generalisation among neutral stimuli during fear condition paradigms, both in healthy controls (Dunsmoor et al., 2013; Laufer & Paz, 2012; Schechtman et al., 2010) and in patients with anxiety disorders (Laufer et al., 2016; Lissek et al., 2009). The generalisation is another example of indirect measure of similarity, because the higher the similarity between stimuli, the wider the generalisation between these stimuli.

1.3.1 Neuroimaging studies

The number of neuroimaging studies in emotional similarity research is limited. To our knowledge, no neuroimaging studies have investigated neural differences in explicit judgments of similarity among the prevalent stimuli in research of emotional cognition, namely, natural, complex neutral and emotional picture scenes. Only a handful of studies have combined behavioural measures of similarity with neural data by using RSA. The results of these studies might help in understanding the brain regions which code the similarity among emotional stimuli. In these studies, during the fMRI scan participants were asked to attend to pictures while performing non-emotional rating tasks (e.g., ratings of indoor versus

outdoor scenes). This was combined with behavioural judgements of similarity among the experimental stimuli. They found that brain activity patterns in regions involved in emotional processing, such as the insula and the ventromedial prefrontal cortex (VMPFC), represent the similarity structure between emotional and neutral stimuli observed at behavioural level (Chavez & Heatherton, 2015; Levine, Wackerle, Rupprecht, & Schwarzbach, 2018).

Additionally, indirect evidence about what make two emotional stimuli similar to each other at neural level is gleaned from neuroimaging investigations of emotional processing and categorisation. These mainly aimed at investigating how the brain codes the relationship between specific emotions, supporting either a categorical (Ekman & Friesen, 1976), a dimensional (Russell & Pratt, 1980), or a constructionist view (Barrett, 2017). In these studies, participants were asked either to passively look at images, to attend to the feelings they evoke, to rate the valence and the arousal of these feelings, or to rate the valence and arousal of the picture and categorise it according to emotional labels (Baucom, Wedell, Wang, Blitzer, & Shinkareva, 2012; den Stock, Vandenbulcke, Sinke, & de Gelder, 2014; Edmiston et al., 2013; Hrybouski et al., 2016; Machajdik & Hanbury, 2010; Motzkin, Philippi, Wolf, Baskaya, & Koenigs, 2015; Ohira et al., 2006; Sakaki, Niki, & Mather, 2012; Yuen et al., 2012). The results of these studies were discrepant, probably because of the different perspectives of emotions adopted and methods used to elicit the emotions (Wager et al., 2015). In particular, locationist studies attempted to discover the unique brain feature associated with each emotional category, by adopting a one (brain region)-to-one (emotion) approach. For example, fear has been consistently localised in the amygdala (LaBar, Gatenby, Gore, LeDoux, & Phelps, 1998; LeDoux, 2007; Öhman, 2009), disgust in the anterior insula (Calder, 2003; Jabbi, Bastiaansen, & Keysers, 2008; Wicker et al., 2003), sadness in the anterior cingulate cortex (F. C. Murphy, Nimmo-Smith, & Lawrence, 2003; Phan, Wager, Taylor, & Liberzon, 2002), anger in the orbitofrontal cortex (F. C. Murphy et al., 2003; Vytal & Hamann, 2010), and happiness in the dorsomedial prefrontal cortex (DMPFC) (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012). As highlighted by Lindquist et al. (2012), supports for a locationist account would be found if instances of an emotion category (e.g., fear) are consistently and specifically associated with increased activity in a brain region (or in a set of regions within a network) across multiple

published studies. However, first, many studies showed that the aforementioned regions are associated with multiple categories of emotions (Lindquist et al., 2012), and during many other sensory, perceptual and cognitive functions (LeDoux, 2012; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). Second, it is not clear whether the findings from the locationist literature are reliable enough or consistent across studies (Wager et al., 2015). For these reasons, a psychological constructionist approach to emotion is preferable. According to this perspective, emotions are 'situated conceptualisations', that is, subjective interpretations of what is happening around us. Emotions arise from the interaction among many brain regions, interconnected in large-scale networks, according to a many-to-one approach. These brain regions are implicated not only in emotional processing, but also in more 'cognitive' functions, such as conceptualization (simulation of previous experiences), language (representation and retrieval of semantic concepts), and executive attention (attention and working memory).

However, this represents only indirect evidence of the neurobiological underpinnings of emotional similarity. The neural mechanism that allows emotion to influence overall perceptions of similarity is still unknown, as are putative neural differences during explicit judgements of similarity between natural, complex neutral and emotional events.

1.3.2 Limitations in emotional similarity literature

Although the emotional similarity literature provided interesting and relevant results, it is also limited in several important ways. First, most studies used decontextualized, simple stimuli, such as emotional faces, or words, a choice that yields more experimental control at the cost of ecological validity. This is particularly important because the known influence of context on emotional categorisation (Barrett, 2017). For example, Avierez et al. (2008) observed this effect in a study about emotional categorisation, where participants were asked to indicate the category that best describes the facial expressions. They were less accurate in categorising sad faces embedded in a fearful than in a sad context: they were more likely to categorise sad faces as fearful when the faces appeared in a fearful context than when they appeared in a sad context. The same effect was observed in the categorisation of disgusted faces embedded in a pride context

(Aviezer et al., 2008). Future studies in emotional similarity should adopt complex stimuli, which depict both emotional and neutral real-world scenes, such as those provided in well-validated datasets, the International Affective Picture System (IAPS) (Lang et al., 2008) and the Nencki Affective Picture System (NAPS) (Marchewka et al., 2014). So far, these more complex stimuli have seldom been used to study emotional similarity (Chavez & Heatherton, 2015; Chikazoe et al., 2014; Gallo et al., 2009; Levine et al., 2018; Talmi & McGarry, 2012).

As implied above, one of the reasons that research on semantic memory and emotional similarity shied away from these more life-like picture scenes might be because there are many factors to control for during the stimuli selection. To mention some of them: the low-level visual measures (e.g., luminance, contrast, and color), the visual complexity of the pictures, the different degrees of similarity among taxonomic levels, the action(s) that the situation can afford, and the thematic similarity within emotional stimuli. In particular, as explained by Talmi and McGarry (2012), emotional stimuli are more thematically inter-related than the neutral stimuli found in validated datasets. For example, the term *car accident* may be related to *hospital*, and then to *death* in a common scenario, while neutral stimuli, such as *architecture*, *telephone* and *laundry*, are less inter-related thematically. In addition, the range of themes within the set of negative and arousing pictures (e.g. death, violence, car accidents, hospital scenes, and assaults) is reduced compared to those within the neutral images. This is also in line with higher ratings of content overlap among arousing (both positive and negative) than neutral IAPS stimuli, observed by Gallo et al. (2009) (Gallo et al., 2009).

To our knowledge, there are no studies which controlled for all these factors, and this represent a further limitation in emotional similarity literature. For example, few recent studies have controlled complex pictures (positive, negative, neutral) for visual properties, as well as for some elements of semantic similarity – animacy (Chikazoe et al., 2014) and social/non-social (Chavez & Heatherton, 2015). However, like other studies (Levine et al., 2018; Yuen et al., 2012), they did not control the stimuli for thematic similarity. In the study by Chavez et al. (2015), the negative categories (i.e., social: ‘depiction of pain’ and ‘people crying’; non-social: ‘polluted water’ and ‘dirty toilet’) look more thematically related compared to neutral (i.e., social: ‘person at a computer’ and ‘person on the phone’; non-social:

‘a stack of book’ and ‘a spoon’) (Chavez & Heatherton, 2015). It is paramount to control for these factors to be able to decouple the effect of emotions and of other factors (e.g., thematic similarity) on the overall perception of similarity, both at behavioural and at neural level. For example, in an unpublished pilot study, we hypothesised higher similarity ratings within 10 negative versus 10 neutral complex pictures, randomly selected from the NAPS database. The results supported our hypothesis. However, we could not conclude whether this effect was related to the emotional nature of the pictures or to a bias in the stimulus selection. This is because we did not control for the higher thematic similarity within the emotional pictures: the range of emotional themes was reduced compared to that in the neutral set. The same reasoning would be valid at the neural level, if we observe higher similarity within the activity patterns in occipitotemporal regions associated with emotional than neutral stimuli. Indeed, without a method to select natural scenes in a way that is representative of their frequency in the environment it is difficult to conclude that emotional stimuli are represented as more similar at neural level than neutral stimuli. To our knowledge, no studies investigated behavioural or neural differences between neutral and emotional complex stimuli during direct similarity judgements.

1.4 Conclusion and future directions

Emotional similarity is a core construct in neuroscience, because it supports many cognitive functions, including categorization, memory, and learning. It is also involved in mechanisms underlying psychiatric conditions, such as anxiety disorders. However, very little is known about what makes us perceive real-life emotional experiences as similar. At the behavioral, or computational, level, most of the studies showed that we implicitly consider the valence and the arousal as relevant dimensions during similarity judgements. Although these studies were very successful in relating behavioral and neural data using innovative MVPA, they mainly used very simple and ‘non-naturalistic’ emotional stimuli.

At the neural, or implementation level, we gleaned indirect evidence about brain regions involved in emotional similarity from research on the structure of the emotional representation of complex stimuli. However, they do not explain which

mechanisms lead to the activity associated with those stimuli. As suggested by Barsalou (2017), this is a common mistake in neuroscience; most studies are related to the computational and the implementation levels. They ignore the algorithmic level, namely the latent mechanisms within the 'system' brain 'that performs the task' (Barsalou, 2017). Future studies should combine all these levels of explanation in MVPA emotional similarity studies, which will benefit of new and well-controlled set of stimuli. This may help in unveiling the influence of emotional similarity on the overall perception of similarity. Finally, we might discover any neural and behavioral differences in perceived similarity between emotional and neutral real-life events, to understand whether emotional similarity is a facet of semantic similarity or a further dimension in a complex semantic space.

2. Chapter: Materials and methods

Measuring similarity is complex because it relies on a series of stimulus characteristics (e.g., visual, semantic, emotional). Estimates of similarity are also affected by inter-individuals differences (e.g., appraisal factors, past and present experiences), and the different experimental contexts (e.g., the set of items in a choice situation) (Martina Riberto et al., 2019). In the last decade, similarity has been studied at behavioural and neural level, using a variety of experimental procedures, ranging from pairwise ratings to multi-arrangements methods. In behavioural tasks, it has been estimated using direct (e.g., similarity judgements) or indirect (e.g., generalisation) measures. Neural similarity is computed as correlation among neural patterns associated with the experimental conditions. Similarity judgements tasks have been mainly conducted using visual stimuli, including datasets of simple (e.g., shapes) and more complex (e.g., scenes) stimuli.

In this chapter, we summarised the experimental procedures and data analyses mostly used in similarity literature. First, we described stimulus sets generally used in similarity tasks, focusing on visual stimuli, and highlighting their confounding factors and limitations (*section 2.1*). Then, we explained the experimental procedures in similarity literature (*section 2.2*), whereby explicit (*section 2.2.1*) and implicit similarity estimates can be derived (*section 2.2.2*). We also defined neural similarity in the context of Representational Similarity Analysis (*section 2.3*), concluding with a summary of the statistical analyses conducted (*section 2.4*).

List of abbreviations

Abbreviations	Meaning
RSA	Representational Similarity Analysis
RDM	Representational Dissimilarity Matrix
CS	Conditioned stimulus
GS	Generalisation stimulus

US	Unconditioned stimulus
fMRI	Functional magnetic resonance imaging
MA	Multi-arrangements
PC	Pavlovian conditioning
ROI	Region of interest
MDS	Multidimensional scaling
VG	Visual generalisation
AG	Auditory generalisation

2.1 Materials

In similarity tasks, materials ranged from simple and visually complex stimuli, including words (Roberts & Wedell, 1994; Talmi & Moscovitch, 2004), shapes (Goldstone et al., 1997), faces (Said et al., 2010; van Tilburg & Igou, 2017), objects (Charest et al., 2014; Kriegeskorte, Mur, Ruff, et al., 2008), and scenes (Greene et al., 2014; Levine et al., 2018). Fewer studies on similarity judgements have been conducted in auditory (Marks, 1987; Sloutsky & Napolitano, 2003), tactile (Gaißert, Bülthoff, & Wallraven, 2011; Marks, 1987) and olfactory (Carrasco & Ridout, 1993; Davis, 1977) sensory modalities.

In the emotional similarity literature, research mainly focused on emotional categorisation and learning. Stimuli were either emotional per se' (e.g., fearful face, image of car accidents) or being emotionally conditioned (e.g., fear, reward). Authors explored how emotional concepts were organised in a bidimensional space, with valence and arousal as orthogonal dimensions (Russell & Bullock, 1985) or how the generalisation gradient changed according to the similarity between different items and the conditioned stimulus (Lissek et al., 2009). Recent studies used similarity-based techniques (e.g., RSA) to investigate how levels of valence (e.g., negative and positive) or basic emotions (e.g., fear, disgust) are encoded into stable neural patterns. They mainly adopted simple emotional stimuli (Liu, Liu, Zheng, Zhao, & Fu, 2021), and fewer complex scenes (Chavez & Heatherton, 2015; Chikazoe et al., 2014), probably because of the variety of

confounding factors (e.g., low-level visual measures, taxonomic and thematic relationships) to control for. One way to take them into account is to match the experimental conditions (e.g., negative and positive emotions) on low-level visual measures, such as luminance, colour, visual complexity, and on semantic similarity. Most of the studies on emotional similarity controlled for visual similarity, omitting the latter. In this thesis, we controlled for both of them, as differences in similarity between emotional and neutral stimuli might be better explained by differences in visual or semantic than emotional features. For example, Madan et al. (2018) observed that highly-arousing pictures were subjectively rated as more visually complex than non-emotional pictures, namely arousal-complexity bias (Madan, Bayer, Gamer, Lonsdorf, & Sommer, 2018). They suggested to control emotional and neutral images using objective (e.g., JPEG compression) rather than subjective measures of visual complexity. Moreover, it is important to note that high-level aspects of scenes (e.g., depicted theme, category) and low-level features (e.g., colour) are inextricably linked (Groen, Silson, & Baker, 2017). For example, the 'red' value on the RGB spectrum of emotional pictures depicting bloody scenes of violence will be higher than for non-emotional pictures. Finally, negative emotional stimuli are more semantically related than randomly-selected neutral stimuli (Talmi, 2013), as they co-occur more frequently than neutral stimuli in the same theme or scenario (e.g.., violence, death). A poor control of semantic similarity might result in a strengthening in similarity among emotional stimuli that are more related to thematic than emotional similarity.

We tested any differences in behavioural and neural similarity measures between emotional and neutral experiences both using simple and visually complex stimuli. In **chapter 3**, we selected 10 neutral and 10 emotional faces, evoking different basic emotions (i.e., disgust and fear) from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist, 1998). This allowed us to control for semantic similarity, as stimuli were all from the same semantic category (e.g., faces) and for apparent visual differences between faces (e.g., gender), as we selected only male faces. In **chapter 4**, we used emotional and neutral real-world photographs from the Nencki Affective Picture System (NAPS) database (Marchewka et al., 2014) and Google Images. These images contain a lot of information beyond the persons themselves, placing them in a rich and realistic context. In experiments involving complex pictures, we instructed participants to focus on the overall

meaning of each picture, discarding irrelevant visual details during the judgements. We created two datasets of negative and neutral scenes, the second of which afforded to control for semantic and visual attributes. For the first dataset (*experiment 1*), we randomly selected scenes from the same semantic category ('people in outdoor situations'), half of them evoke negative emotions and half were neutral. Emotional and neutral pictures were comparable on low-level visual measures. However, the range of emotional themes was reduced compared to that in the neutral set. Thus, in the second dataset (*experiments 2-3*), we controlled for difference in thematic similarity, by selecting natural scenes in a way that all the categories depicted realistic events that do not co-occur in the environment. It consisted of 72 real-world colour photographs, which represented one or more people in outdoor situations. We divided them into 4 categories according to the scene that was depicted, resulting in 18 images per category. Two of the categories were neutral, and two were emotionally-arousing and negatively valenced, as revealed by valence and arousal ratings provided by an independent sample of participants. These latter categories represented either poverty scenes (Emotional category 1, E1) or car accidents (Emotional category 2, E2). The neutral categories portrayed either people talking on the phone (Neutral category 1, N1) or hanging laundry to dry (Neutral category 2, N2). The full set of pictures can be found in Figure 2.1.

Figure 2.1. Second database of complex pictures, divided into four categories (18 pictures within each category), two of them are negative emotional and two neutral. The first neutral category (N1) represents people talking on the phone and the second one (N2) people hanging the laundry. The first emotional category (E1) depicts poverty scenes, the second one (E2) car accidents. The full set of pictures can be found at https://dtalmi.wixsite.com/website/resources

Finally, in **chapter 5**, we opted for conditioning simple stimuli, ensuring an optimal control of visual and semantic similarity, and for this specific semantic category (i.e., animals). We adopted visual and auditory stimuli, as we focused on similarity across sensory modalities in an aversive conditioning paradigm. Specifically, in *experiment 1*, we selected 80 images and 80 vocalisations of animals (i.e., 'mammals' and 'birds'), which belonged to 8 basic-level categories (e.g., cow, sparrow). Four of them served as CS, and the remaining four categories as GS. Each basic-level category consisted of 10 images and 10 sounds of different exemplars (e.g., different breeds). As *experiment 2* involved an fMRI task that required a rich conditions design (Nili et al., 2014), we doubled the number of both visual and auditory stimuli. Images were selected in Google Images, and vocalisations from www.soundsnap.com, www.epidemicsound.com and publicly available resources on the internet.

2.2 Experimental paradigms

2.2.1 Explicit similarity judgements tasks

In similarity judgement tasks, participants are asked to judge the stimuli according to which features they share. Stimuli are presented in triads, or in pairs, or around the perimeter of a circular 'arena'. In forced-choice triad tasks (Jackson et al., 2015; Niedenthal, Halberstadt, & Innes-Ker, 1999; Simmons & Estes, 2008), three stimuli (one target and two options) are displayed together, and participants choose the option more similar to the target.

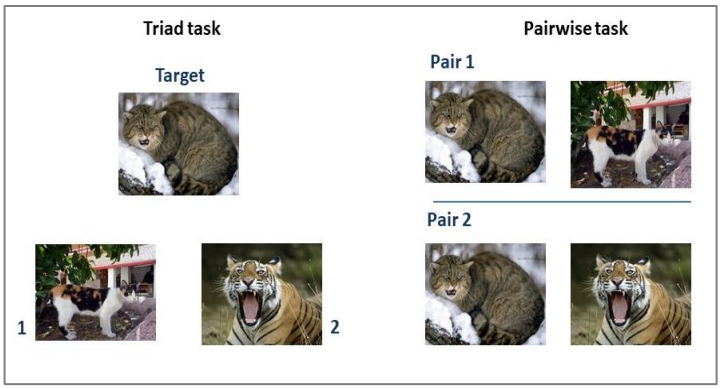


Figure 2.2. Example of the triad (left) and pairwise task. Triad tasks are forced-choice similarity judgement tasks, because participants have to choose which stimulus (1 or 2) is more similar to the target. In this situation, the similarity between the target and 2 is reduced, because of the presence of stimulus 1. In pairwise tasks, each pair is rated independently. This allows participants to consider also differences in similarity between categories.

For example, in the triad presented in Figure 2.2, participants would choose option 1, because both depict cats. However, they would underestimate the similarity between categories, as the target is also similar to 2, because both are feline. Conversely, in pairwise tasks (Pollak, Cicchetti, Hornung, & Reed, 2000; Russell & Bullock, 1985; Wisniewski & Bassok, 1999), each pair is independently rated, generally on a 7-point scale. We opted for the pairwise than triad presentation, because participants can also focus on differences in similarity between categories. However, one of the limits of pairwise similarity judgement tasks is the relatively long duration, and thus it is feasible for stimulus sets with a small number of items (Kriegeskorte & Mur, 2012).

An alternative task in experiments with relatively large number of stimuli is the MA task. Participants have unlimited time to drag and drop the stimuli in a low-dimensional space (circular 'arena') according to their similarity, such that similar stimuli were placed close to each other and dissimilar stimuli apart. In each trial, participants focused on a specific subset of stimuli, which are usually positioned along the perimeter of the arena. A trial ended when participants arrange all the stimuli in the arena. The task is concluded when participants judged all the

possible combinations. Kriegeskorte and Mur (2012) also showed the high test-retest reliability (r=0.81) as well as external validity with pairwise tasks (Kriegeskorte & Mur, 2012). One of the drawbacks of the MA task, especially in emotional similarity research, is the reduction of emotion differentiation to locations in a bidimensional valence and arousal space (Grandjean, Sander, & Scherer, 2008). In our experiments, part of this was taken into account by testing the same research question using different similarity tasks.

We opted for the pairwise rating task in experiments with a small number of stimuli **(chapter 3**; **chapter 4:** *experiment 1*). Also, *experiment 2* in **chapter 4** required ratings of paired complex images. However, because of the high number of stimuli, we divided the 72 pictures into two subsets ('even' and 'odd', n=36 within each subsets) and we focused on the pairs of interest (170 total pairs). Conversely, in the other experiments that involved a rich stimulus design for fMRI tasks, we investigated similarity perception using MA task. In particular, in **chapter 4** (*experiment 3*), after a visual complexity rating fMRI task, participants arranged in the similarity space subsets of complex pictures. The total duration of the task was approximately 50 minutes. Finally, in **chapter 5** (*experiment 1*), participants performed the MA task before and after the aversive conditioning paradigm. Because of time constraint, we divided the images into 'even' and 'odd' and each subject was randomly assigned to one of the subsets. In order for participants to focus on the similarity both within and between basic-level categories, we opted for splitting the images into 3 different arenas, one for mammals, the other for birds, and one 'mixed' arena with few exemplars of both mammals and birds. The purpose of the mixed arena was to compare within-category and between-category similarity. The task ended after approximately 20 minutes.

In these tasks, similarity is estimated directly, because participants are asked to judge explicitly the similarity among stimuli. As our main research question concerned differences in similarity perception between emotional and neutral stimuli, we choose direct measures of similarity for most of the experiments. Specifically, in **chapter 3 and 4** (*experiments 1-2*) we used the similarity ratings, and the Euclidean distance between stimuli in the arena in the MA task in **chapter 4** (*experiment 3)* and **chapter 5** (*experiment 1)*. However, implicit measures of similarity can be derived from these tasks, including the reaction times (RTs), reasoning that participants are quicker in judging the similarity between stimuli that

share very few common aspects (Chen et al., 2013; Iordan et al., 2015) and the semantic relatedness, which refers to the degree to which two concepts or words are related in semantic memory. The more semantically related two items are, the more similar they are (Talmi & Moscovitch, 2004). Finally, participants' accuracy in categorisation and generalisation are further implicit similarity measures. This is because stimuli in the same category are more similar than stimuli from different category (Greene et al., 2014), and the higher the similarity between stimuli, the wider the generalisation between them (Dunsmoor, Prince, Murty, Kragel, & LaBar, 2011).

2.2.2 Implicit similarity tasks: aversive conditioning paradigms

In **chapter 5**, we also included implicit measures of similarity, that is, participants' accuracy in categorising the CSs into CS+ and CS- and in generalising this learned associations to similar GSs. We selected this specific implicit similarity measure, because using the semantic relatedness between stimuli would have biased participants to focus on the semantic similarity, and the RTs would allow us to pick up only big differences in similarity between stimuli. We used aversive conditioning to induce an emotional impact on images of animals and to explore its effect on the similarity among stimuli that resemble the CS+. We instructed participants that their goal was to understand which images and sounds of animals predicted the US (loss vs the saving of 2 Israeli shekels, NIS). In case of a loss, the money was taken from the initial amount that they received before starting the experiment. Participants learned the association in a series of PC and GS trials. The former included visual stimuli only, whereas the latter both visual (in VG) and auditory (in AG) stimuli. During PC trials, images of each CS were displayed on a blank screen. Participants rated whether they predicted either the loss or the saving of 2 NIS (i.e., lose vs saving) and their degree of confidence (i.e., sure vs almost sure). Then, according to the nature of the CS, a feedback appeared on the screen (CS+: '*you lost 2 NIS*'; CS- '*you saved 2 NIS*'). As previously mentioned, GS trials were identical to PC trials, except that the US was not shown. While participants were conditioned using visual stimuli, generalisation trials included either visual or auditory stimuli. The purpose of PC trials was to enable learning, but otherwise, they were not the focus of our main hypotheses, and they

were therefore fewer in number than GS trials. We were interested in neural measures of similarity between threatening and not threatening GS, which were computed in the context of Multi-voxels pattern analysis (MVPA) techniques, such as Representational similarity analysis (RSA).

2.3 Representational similarity analysis (RSA)

In the last decade, similarity-based techniques have been used to investigate how stimuli from different categories are encoded in stable neural representations in specific brain regions. This assumes that high neural similarity corresponds to high correlation among neural activity patterns associated with the experimental stimuli. Among these techniques, RSA has been gaining popularity in the neuroimaging literature. It can be conceived as a hub that relates findings from the three major branches of neuroscientific research (e.g., behavioural, neural and computational data) by computing the correlation among them (Kriegeskorte, Mur, et al., 2008b). RSA allows for establishing a second-order isomorphism between behavioural and neural data, that is, unveiling the brain regions involved in representing inter-stimulus relationships observed at behavioural level, by computing the correlation among them. We will refer to this as 'brain-behaviour correlation'. In case of computational data, it is also possible to map which brain regions carried out the computations simulated in the model, through correlations between neural patterns and the computational model (brain-model correlations). Alternatively, we can investigate how different experimental conditions are encoded in specific ROIs, and the differences among their neural representations, corresponding to differences in neural similarity. We will refer to this as 'ROIs RSA'. Using ROIs RSA, we can also compare how different stimuli are encoded in coherent neural patterns across sensory modalities. Finally, adopting a functional coupling perspective, it is also possible to explore neural encoding across brain regions, an approach named 'representational connectivity' (Nili et al., 2014). This allows to explore the neural similarity across brain regions that are functionally connected.

As shown in Figure 2.3, after conventional fMRI preprocessing, the first step requires the estimation of a RDM for each data type that indicates the degree to which each pair of conditions is distinguished. It is an *n* x *n* matrix, wherein the row and the columns represented *n* experimental stimuli, and each cell a dissimilarity

measure between stimuli in each pair. Dissimilarity measures included correlational distance (1 minus the correlation between patterns), which ranges from 0 to 2 (0 for perfect correlation, 1 for no correlation, and 2 for perfect anticorrelation) (Haxby et al., 2001). Alternative measures are the Euclidean distance (Edelman, 1998), the Mahalanobis distance (Kriegeskorte, Goebel, & Bandettini, 2006) and the absolute value of the regional-average activation difference in fMRI analysis. The latter is sensitive only to the overall level of activation, whereas the correlation distance (1 − correlation) normalizes for both the overall activation (which could be attributed e.g., to attention) and the variability of activity across space. The Euclidean distance combines sensitivity to pattern shape, spatial-mean activity level, and variability across space. Using the Euclidean distance yields an RDM resembling both the one obtained with correlation distance and the one obtained with absolute activation difference.

In light of these considerations, in fMRI experiments we chose the rank-correlation distance as measure of neural dissimilarity (1- Spearman's correlation), as it is not sensitive to the global activity level and we did not wish to assume a normal distribution underlying dissimilarity estimates (Kriegeskorte, Mur, et al., 2008b).

As depicted in Figure 2.3, in a RDM on behavioural data, the dissimilarity measures consisted either in standardizing the similarity ratings and subtracting it from 1 (to obtain a dissimilarity estimate), or the Euclidean distance between stimuli in the low-dimensional space. In a neural RDM, the correlational distance is computed as correlation across betas (or t values) of all the voxels in a ROI associated with the stimuli in each pair. An alternative is using the 'searchlight RSA', a very precise localisation technique, wherein 3 × 3 × 3 voxels spherical cluster is moved throughout the brain and at each location a correlational distance (among betas or t values) is assigned to the central voxel of the sphere. This measure quantified the dissimilarity across voxels in a given searchlight sphere for each specific pair. As a consequence, each RDM was symmetrical about a diagonal of zeros that represented the dissimilarity of each stimulus with itself. The only exception to this is an RDM resulting from correlation between RDMs. This is the case with RDM across ROIs, sensory modalities (e.g., visual and auditory), data type (e.g., behavioural and neural), time points (e.g., before and after aversive conditioning), as the diagonal indicates the dissimilarity of the same stimulus between the different experimental conditions.

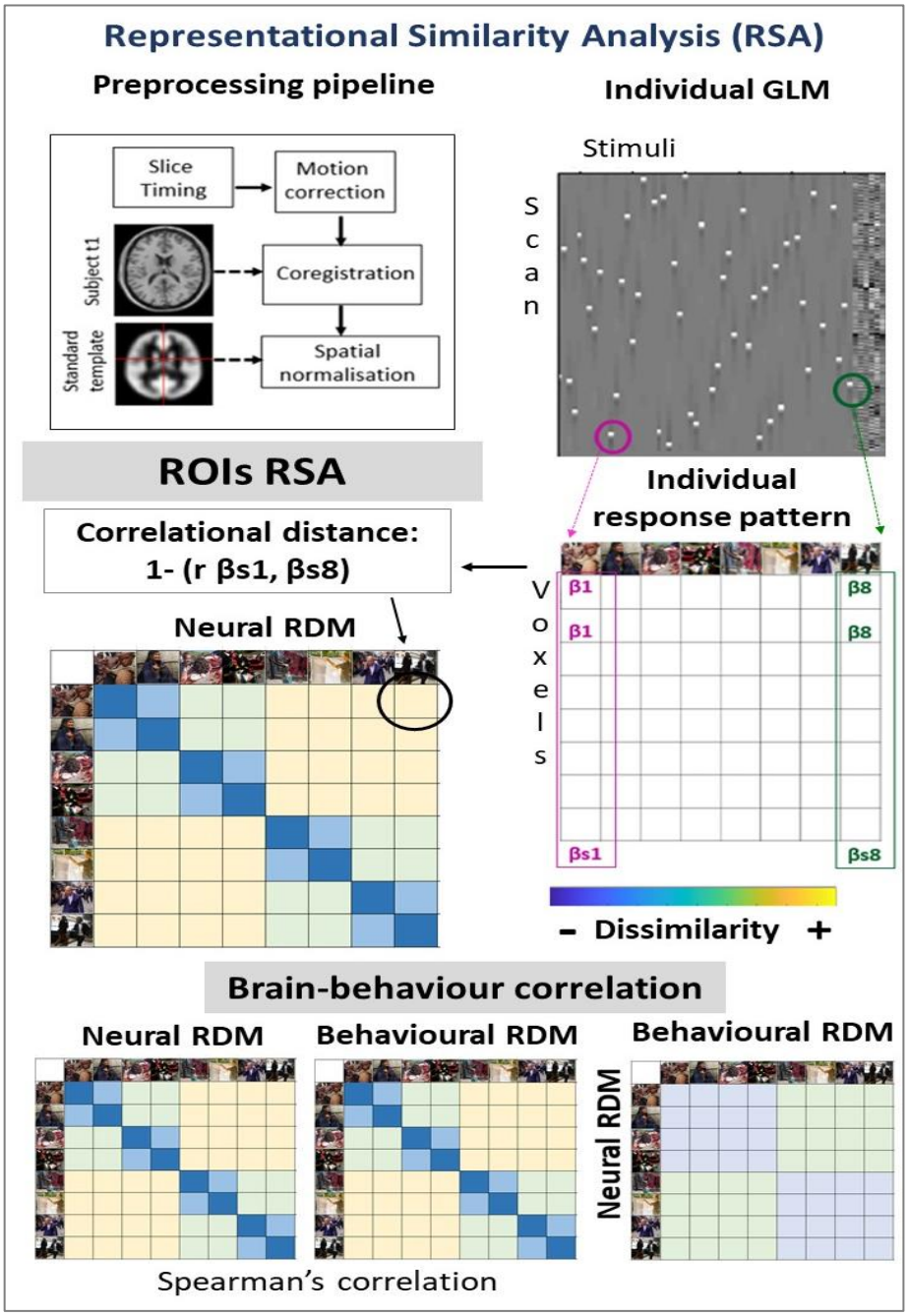


Figure 2.3. Graphical description of the different processing steps in the RSA framework. After a conventional temporal and spatial preprocessing ('preprocessing pipeline'), normalised images from each voxel were analysed using the general linear model (GLM). Each stimulus was modelled as a separate event beginning with stimulus presentation onset, and included in the model as regressor of interest ('individual GLM'). From this GLM analysis, we obtained a single beta image for each stimulus within each voxel ('individual response pattern'). Next, we computed the correlational distance (1-Spearman's correlation) across betas of all the voxels in a ROI associated with the stimuli in each pair. These represented the entries of an n x n neural RDM, wherein the rows and the columns are the experimental stimuli ('neural RDM'). This is symmetrical about a diagonal of zeros that represented the dissimilarity of each stimulus with itself. Blue colours denote low dissimilarity (high similarity). Other than investigating differences in neural dissimilarity among experimental conditions ('ROI RSA'), it is also possible to combine neural and behavioural RDMs computing the Spearman's correlation among them, and then convert it into correlational distance ('Brain-behaviour correlation'). This results in n x n RDM, wherein the rows and the columns indicate the behavioural and neural data, respectively, associated with each stimulus, and each cell the dissimilarity between neural and behavioural data. This RDM is not symmetrical, as the diagonal indicates the dissimilarity of the same stimulus between the different neural and behavioural data.

In all our experiments, we used RSA to estimate measures of behavioural and neural similarity, and to establish a second-order isomorphism between them. Specifically, in experiments involving pairwise ratings task (i.e., **chapter 3**; **chapter 4:** *experiments 1-2*), we first standardized the similarity ratings, by subtracting 1 (the lowest similarity rating) from each rating *x*, and then divided by 6 (highest similarity rating - lowest similarity rating), and then we transformed them into dissimilarities, by subtracting the ratings from 1. In experiments with MA task (**chapter 4:** *experiment 3;* **chapter 5:** *experiment 1*), we used the Euclidean distance as dissimilarity measure. These distances represented the entries in the behavioural RDMs. In **chapter 4:** *experiment 3*, we combined ROIs RSA and a brain-behaviour correlation approach to unveil which brain regions represented participants' similarity space (brain-behaviour correlations), and whether the neural representations of emotional and neutral stimuli in these regions differed (ROIs RSA). We adopted the searchlight RSA to compute the neural RDMs. Finally, in **chapter 5:** *experiment 2*, we explored differences in neural similarity between threatening and not threatening stimuli, within and across sensory modalities, adopting a ROIs RSA approach. Data from both the neural and behavioural RDMs are then subjected to different statistical analyses, wherein the mean and the standard deviation of the conditions of interest are extracted from each RDM, and then compared at group-level according to the research questions.

2.4 Statistical data analysis

2.4.1 Similarity measures

In all our experiments, for each participant, we extracted the mean and the standard deviations of the conditions of interest from each RDM to investigate differences in behavioural and neural similarity between emotional and not emotional stimuli. In case of neural data, this procedure is performed in each ROI separately. In **chapter 3**, we compared the similarity among faces expressing different emotions (i.e., fear and disgust) with the similarity of visually different neutral faces (NN), and visually identical faces (Identity), which differed on emotionality (i.e., one of them is emotional and the other one is neutral). In **chapter 4** (*experiment 1)*, we compared the similarity between emotional (EE) and neutral

(NN) stimuli, and in **chapter 4** (*experiments 2-3)* between the two emotional (E12) and the two neutral (N12) categories. The latter was also performed at neural level in **chapter 4** (*experiment 3)*. Finally, in **chapter 5** (*experiment 1)*, we tested differences in similarity perception after aversive conditioning between threatening (i.e., CS+ and GS+) and not-threatening (i.e., CS- and GS-), and in **chapter 5** (*experiment 2)* differences in neural similarity between GS that were semantically related to the CS+ (GS+), with those that belonged to another semantic category (GS-). The means of each condition were then entered in paired t-tests or repeated-measures ANOVAs (according to the number of levels of each within-subject factor), testing the null hypothesis of no differences in (behavioural or neural) similarity between emotional and non- emotional conditions. Bonferroni post hoc corrections for multiple comparison are used to explore the nature of the effects.

In brain-behaviour correlation analyses, the similarity between brain and behavioural RDMs for each stimulus was estimated using pairwise Spearman's correlations. This provides a correlational map between the behavioural and the brain RDMs for each subject, which reveals where the similarity space is best represented in the brain (highest correlation). The correlational coefficients were Fisher's z transformed, and inference was performed at each voxel by performing a one side signed rank test across subjects, testing the null hypothesis of no correlation between brain and behaviour RDMs. The resulting p values (uncorrected) were thresholded to control the false-discovery rate (FDR).

In order to visualize how the stimuli are displaced in a low-dimensional space according to their similarity, we performed MDS analysis. When rendered in the space by MDS procedure, the similarity among stimuli map the cognitive structure of those stimuli (Barrett, 2004). We performed it on the behavioural data, although it is extendible to any similarity measures. Specifically, we entered as input the similarity estimates amongst stimuli, and the output is a geometric space, wherein proximities reflect the similarity among stimuli. This space is defined by a number of dimensions (or axes), and the stimuli are represented by their coordinates on these dimensions. The dimensions characterise implicit components of the cognitive structure of the experimental stimuli. The fit of the MDS solution is typically determined by an iterative process, in which a goodness-of-fit measure (*Stress*) is estimated (Shinkareva, Wang, & Wedell, 2013). This measure

represents how far the observed data are from the predicted data, therefore, low values demonstrate a better fit of the MDS representation. Stress is closely related to the number of dimensions included in the model: it increases with fewer dimensions, but using fewer dimensions could result in a distortion of the true structure. Conversely, Stress decreases with many dimensions, whereas this solution could lead to an overfitting of the data. A Stress value below 0.15 reflects an adequate fit and below 0.10 is an excellent fit (Shinkareva et al., 2013).

2.4.2 Aversive conditioning (i.e., accuracy, latency and pupil diameter) and emotionality ratings

In **chapter 5**, indirect measures of similarity (i.e., accuracy in generalisation of loss expectancy) and latency were used as measures of aversive conditioning. We calculated accuracy scores by dividing the number of correct answers, that is, when participants rated they were either sure or almost sure to lose money in a CS+/GS+ trials (and vice versa for CS-/GS-), by the number of stimuli within each condition. Accuracy was measured separately for PC and GS trials. Successful learning was defined as an accuracy score higher than chance level (50%) across CSs during PC trials. In a similar manner, we defined successful generalisation as above-chance accuracy, computed separately for visual and auditory modalities. We analysed learning and generalisation with three one sample t-tests (separately for PC, VG and AG), testing the null hypothesis of no difference between the average accuracy and the chance level (50%). We also expected increased accuracy scores and decreased RTs over time in each of these three conditions. We tested this hypothesis by entering average accuracy and latency within the first and the second half of the task in three repeated-measures ANOVAs (separately for PC, VG and AG, one for accuracy and one for latency), with time as within-subject factors.

Additional measures of successful conditioning included neurophysiological indicator of arousal, such as the pupil diameter (PD). In **chapter 5** (*experiment 2)*, we recorded it using eye-tracking as manipulation check of aversive conditioning, expecting wider PD associated with threatening than not-threatening stimuli. We predicted that this effect might follow trial onset, response time, and US time. In order to test this, we estimated average PD for our conditions of interest,

separately for each sensory modality, that is, PD GS+ (onset), PD GS+ (response), PD GS- (onset), and PD GS- (response). The same conditions were valid in PC trials, with the addition of PD when US was delivered (i.e., PD CS+ (US) and PD CS- (US)). We entered them as dependent variables in different paired t-tests with stimulus type as within-subject factor, one for each segment (separately for PC, VG, AG).

Finally, in **chapter 4** we estimated average valence and arousal ratings for emotional and neutral stimuli, and **chapter 5** (*experiment 2)* for CS+, CS-, GS+ and GS-, following the procedure suggested by Lang et al. (2008) (Lang et al., 2008). Participants viewed one of images presented in the centre of the screen, and rated each picture on two 9-points scale (valence scale: 1, negative emotions; 9, positive emotions; 5 neutrals. Arousal scale: 1, relaxed; 9, aroused; 5 neutral). We instructed participants to respond as quickly as possible by clicking the appropriate number key, and informed them that there was not a right or wrong answer. We considered emotional pictures as rated less than 4 in the valence scale (negative valence) and more than 6 in the arousal scale (high arousal), whereas the neutral images ranged from 4 to 6 in both dimensions. Valence and arousal ratings were entered as dependent variables in paired t-tests or repeated-measures ANOVAs (according to the number of levels of each within-subject factors), separately for valence and arousal, testing the null hypothesis of no differences in these dimensions between emotional and not emotional conditions.

3. Chapter: Symmetry in emotional and visual similarity between neutral and negative faces

This chapter has been published in *Symmetry* (Martina Riberto, Talmi, & Pobric, 2021).

Abstract

Is Mr Hyde more similar to his alter ego Dr Jekyll, because of their physical identity, or to Jack the Ripper, because both evoke fear and loathing? The relative weight of emotional and visual dimensions in similarity judgements is still unclear. We expected an asymmetric effect of these dimensions on similarity perception, such that faces that express the same or similar feeling are judged as more similar than different emotional expressions of same person. We selected 10 male faces posing different expressions. Each male posed one neutral expression and one emotional expression (5 disgust, 5 fear). We paired these expressions, resulting in 190 pairs, which differed either in emotional expressions, physical identity, or both. Twenty healthy participants rated the similarity of paired faces on a 7-points scale. We report a symmetric effect of emotional expression and identity on similarity judgements, suggesting that people may perceive Mr Hyde to be just as similar to Dr Jekyll (identity) as to Jack the Ripper (emotion). We also observed that emotional mismatch decreased perceived similarity, suggesting that emotions play a prominent role in similarity judgements. From an evolutionary perspective, poor discrimination between emotional stimuli might endanger the individual.

3.1 Introduction

Emotional similarity refers to the tendency to group stimuli together because they evoke the same feelings in us, even when they are visually different. For example, we may judge two different individuals with fearful faces either as similar, because they both express negative emotion, or different, because visually they do not look alike. At present, it is not clear whether different stimulus attributes (i.e., emotional expression, visual features) have a symmetrical or asymmetrical influence on similarity perception. In other words, is Mr Hyde more similar to Dr Jekyll, because they have the same facial features (same identity), or to Jack the Ripper, because of the emotions they trigger in witnesses of their crimes?

The investigation of emotional similarity has a long tradition, with both replicated and controversial results. First, as framed by Russell's circumplex model, participants rate the similarity between emotional stimuli according to their resemblance in valence and arousal. These orthogonal dimensions (valence and arousal) define participants' emotional similarity space, wherein proximities reflect the similarity among stimuli (Russell & Pratt, 1980). This was replicated both in adults and children (Hoemann, Xu, & Barrett, 2019; Russell & Bullock, 1985; Tseng et al., 2014), using simple stimuli, such as words (Cowen & Keltner, 2017; Koch et al., 2016; Talmi & Moscovitch, 2004), objects (Biondi, Franzoni, Li, & Milani, 2016; Leclerc & Kensinger, 2008), and faces (Aviezer et al., 2008; Jamin Brett Halberstadt & Niedenthal, 1997; Mondloch, Nelson, & Horner, 2013; van Tilburg & Igou, 2017), and with more complex stimuli, such as real world photographs (Chavez & Heatherton, 2015; Gallo et al., 2009; Levine et al., 2018). Based on this line of research, an increasing number of studies aim to decode the nature of emotions in the brain (Kragel & LaBar, 2016), particularly where and how valence and arousal are represented, by computing the correlation between behavioural and neural measures of similarity (Chikazoe et al., 2014; King et al., 2019; Todd, Miskovic, Chikazoe, & Anderson, 2020; Yuen et al., 2012).

One of the most controversial findings in the emotional similarity literature is related to asymmetries in similarity judgements between different levels of valence (i.e., negative vs positive). Specifically, in a series of experiments Koch et al. (2016) demonstrated that 'good is more alike than bad', that is, there is higher similarity among positive than negative emotional stimuli (Koch et al., 2016;

Mondloch et al., 2013). By contrast, others report higher semantic relatedness among negative than randomly-selected non-emotional pictures (Talmi, 2013) and wider generalisation in conditioned than unconditioned stimuli in healthy controls (Starita, Kroes, Davachi, Phelps, & Dunsmoor, 2019). One of the reasons for these mixed results might be related to differences in semantic similarity among the various levels of valence of the experimental stimuli used. This may confound the relationship between emotional dimensions and perceived similarity (Martina Riberto et al., 2019). One way to control for this confounding factor is to select simple stimuli, possibly from the same semantic category, such as faces.

Many datasets of prototypical emotional and neutral faces are available to date (Ekman & Friesen, 1976; Lundqvist, 1998; Mollahosseini, Hasani, & Mahoor, 2017; Tottenham et al., 2009). These are widely used in emotion cognition research to uncover how facial expressions are processed and perceived. In general, evidence from neural data shows that regions in the occipitotemporal lobe, including the fusiform face area, the inferior temporal cortex, and the superior temporal sulcus, encode facial identity and similarity among facial expressions (Charest et al., 2014; Guntupalli et al., 2016; Haxby et al., 2011). In addition, Said et al (2010) observed a positive correlation between neural similarity in the posterior superior temporal sulcus and affect-based similarity ratings (Said et al., 2010). In behaviour, faces depicting basic emotions, which share the same valence and arousal, elicited similar subjective experiences in healthy participants (R. E. Plutchik & Conte, 1997; Posner, Russell, & Peterson, 2005). Among basic emotions, happiness is the one mostly recognized with high accuracy and low ambiguity (Leppänen & Hietanen, 2004; Palermo & Coltheart, 2004). Anger and disgust (Pochedly, Widen, & Russell, 2012), as well as fear and surprise (Palermo & Coltheart, 2004) are most frequently confused, probably because of an perceptual overlap, with lowered eyebrows in anger and disgust, and raised eyebrows in fear and surprise (Matsumoto & Ekman, 2008). This similarity in emotional expression and physical appearance might explain part of the overall similarity observed between faces expressing different emotions in the face similarity space (Valentine, Lewis, & Hills, 2016). This is in line with the results from Said et al. (2010) who instructed two groups of participants to rate either the visual or the emotional similarity among faces, and reported high correlation

(r=0.93) between the visual and the affect-based similarity ratings (Said et al., 2010).

However, these studies have so far failed to investigate the relative weight of emotional expression and visual identity in global similarity judgements among faces, since they did not ask participants to focus on one of these features. Only a handful of studies (Jamin Brett Halberstadt & Niedenthal, 1997; A. Nestor, Plaut, & Behrmann, 2016; Valentine et al., 2016; Wegrzyn, Vogt, Kireclioglu, Schneider, & Kissler, 2017) explored the latter effect. Among them, Wegrzyn et al. (2017) asked participants to recognize emotions from faces that depicted two identities (one male and one female), which expressed 7 different emotions. Faces were masked by a grid of white tiles, which started with one tile randomly shown and subsequently one additional tile was revealed every second. Participants were instructed to click a button below the image when they recognized the facial expression, and to select the labelled button corresponding to it in a forced-choice decision task. The Multidimensional scaling (MDS) analysis of the emotions recognition task revealed that faces clustered according to the emotion they expressed in similarity space. Conversely, the MDS with the low-level visual features (light intensity in each pixel) of the faces as input showed that faces were displaced according to the identity they depicted. However, in this study participants were not asked to process inter-stimulus relationships. Conversely, Halberstadt and Niedenthal (1997) manipulated emotions by instructing participants to watch either emotional (positive or negative) or neutral movies, and then to judge the similarity among faces. Participants in the emotional compared to the non-emotional states weighted the emotional dimension of faces more than the gender or head orientation. Taken together, these studies suggest that the relevance of emotional expression and identity may be malleable according to task instructions, and that both are salient features that define participants' face similarity space. However, it seems that these dimensions interact during subjective similarity judgement tasks.

One promising technique to disentangle emotional and visual facial features is to compute objective measures of visual similarity among faces, as in the eigenfaces method (Valentine et al., 2016). According to this approach, the visual similarity among faces is conceptualized as correlation between pixel values of light intensity; the eigenfaces are extracted by performing principal component analysis

(PCA) on the correlations among faces, and represent unique visual features of a set of human faces as dimensions, which define the face-space (Sirovich & Kirby, 1987). This approach has been widely adopted in the context of face recognition and identification, because of the speed of recognition and a higher success rate in comparison to other computational methods (üge Çarıkçı & Özen, 2012). Several studies (Hsu, Tseng, Kang, & Wang, 2013; Sharma et al., 2018; üge Çarıkçı & Özen, 2012; Yuan, Mcdonough, You, & Luo, 2013) used eigenfaces to predict the emotions evoked from images. Success is greater when this method is used, compared to low-level visual features (e.g., GIST, colour histograms). For example, Yuan et al. (2013) developed a novel algorithm based on eigenfaces, Sentribute, which reached a level of accuracy of 82% in predicting image sentiments based on mid-level attributes (Yuan et al., 2013). A similar approach was adopted in another study by Hsu (2013), wherein authors automatically identified and discriminated emotions according to the two-dimensional subspace of valence-arousal (Hsu et al., 2013).

We computed objective measures of visual similarity, in order to control for visual similarity as confounding factor of the effect of interest: asymmetry between emotional expression and identity features on similarity judgements. In particular, we expected that paired faces with different identity that express the same or similar emotions (Mr Hyde and Jack the Ripper) would be perceived as more similar than faces with the same identity, but different emotional expressions (Mr Hyde vs Dr Jekyll), as showed in Figure 3.1. With this aim, we selected negative and neutral faces, which differed in either emotional or visual aspects. We also expected higher similarity ratings for faces with the same emotional expression or same identity (similarity within-category) than for faces with different emotional expressions and identities (similarity between categories). The first prediction represents our main hypothesis; the second one serves as manipulation check, since a good category boundary simultaneously maximizes the within-category similarity, and minimizes the between categories similarity.

3.2 Materials and methods

Participants

A total of twenty healthy participants (13 females, 7 males; mean age 32.10 ± 10.17) were recruited from the University of Manchester to take part in the study. This sample size is comparable to other publications on this topic (Gray, Adams, Hedger, Newton, & Garner, 2013; Leal, Tighe, & Yassa, 2014). All participants had normal or corrected-to-normal vision, and were older than 18 years. Participants provided informed consent prior to the experiment and were reimbursed for their participation. The exclusion criteria were: a history of neurological (e.g., head injury or concussion) or psychiatric (e.g., depression, anxiety) conditions, drug or alcohol abuse, or regular medication that could influence emotional processing. The study was approved by the ethics board number 2018-3619-5928 of the University of Manchester.

Stimuli

Twenty images of faces (562 x 762 pixels) were selected from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist, 1998), which comprises 490 colour pictures of human facial expressions from 70 selected individuals (35 women and 35 men), each displaying six basic emotions (angry, fearful, disgusted, happy, sad, and surprised) and a neutral facial expression. Each expression is photographed from the front. In particular, we selected 10 emotional (five disgust, IDs: 02, 06, 10, 17, 27; five images of fear, IDs: 04, 08, 11, 23 28) male facial expressions, and their neutral equivalents (n=10), which corresponded to the same IDs. We chose fear and disgust to have one emotion that is expressed more on the 'upper face' (fear) and one on the 'lower face' (disgust). We chose this to minimise the visual similarity between emotional faces, such that their similarities were more related to emotional features. Males were selected in order to exclude gender as an additional dimension to consider in the judgement of similarity, which is beyond the scope of this experiment.

Experimental procedure

Participants viewed all possible pairs of the 20 images, resulting in 190 different combinations, presented side by side on a blank screen. Participants were instructed to rate the similarity of each pair by using a 7-point scale (1= low similarity, 7= high similarity). Each trial started with a central fixation cross for 500 ms, the task cue ('how similar do you think these pictures are?') was presented at

the top of the screen, and the judgement scale at the bottom. Participants were told to respond as quickly as possible by clicking the appropriate number key, and were informed that there was not a right or wrong answer. The task ended after approximately twenty minutes.

Data analysis

Similarity ratings. We analysed the similarity ratings using Representational Similarity Analysis (RSA) (N. Kriegeskorte, M. Mur, & P. Bandettini, 2008a), implemented in Matlab R2018, and SPSS. A graphical representation of the conditions of interest and key hypotheses is shown in Figure 3.1. Specifically, the similarity ratings were entered into a 20 x 20 similarity matrix for each participant. The rows and the columns represent the experimental stimuli, and each cell reflects the similarity rating for each pair. Then, for each subject, a Representational Dissimilarity Matrix (RDM) was computed. We first normalized the similarity ratings, by subtracting 1 (the lowest similarity rating) from each rating *x*, and then dividing by 6 (highest similarity rating - lowest similarity rating). Second, we transformed them into correlational distances, by subtracting the ratings from 1. These values were entered into each cell of the RDM. The RDM is therefore symmetric about a diagonal of zeros. Next, we extracted from the single-subject RDM the mean dissimilarities and standard deviations of our conditions of interest, shown in Figure 3.1: within emotional faces (EE), calculated by averaging the dissimilarity within disgusted (EE_D) and within fearful (EE_F) faces; within neutral faces (NN); between emotional and neutral faces with the same identity (ID); between emotional and neutral faces with different identities (EN). The latter served as measure of dissimilarity between categories, the first three as within-category dissimilarity. We also considered the dissimilarity between fearful and disgusted faces (EE_DF) as part of dissimilarity within-category, because the faces in this condition shared negative valence and high arousal. We included this measure to further test our main hypothesis with a dimensional approach to emotions. The dissimilarity measures were entered as dependent variables in two one-way repeated- measures ANOVAs, with conditions as grouping factor. The main hypothesis was tested in the first ANOVA, which included the conditions EE, NN, EE_DF, and ID, and used a planned contrast to test lower similarity (higher dissimilarity) in ID compared to the other conditions, as displayed in Figure 3.1.

The second ANOVA used a planned contrast to test lower similarity (higher dissimilarity) in EN than in EE, NN, EE_DF and ID. Bonferroni post hoc corrections for multiple comparisons ($p<0.05$) were used to explore the nature of the effect.

Multidimensional Scaling (MDS) analysis. In order to visualize the structure of the similarity space, we performed a Multidimensional Scaling (MDS) analysis on the similarity ratings, where proximities reflect similarities among stimuli and are measured on an ordinal scale. The rank order of proximities determines the dimensionality of the space and the metric configuration of the points representing the stimuli (Shinkareva et al., 2013). In line with previous studies in this research field, we assumed this space to be two-dimensional, with valence and arousal as orthogonal dimensions (Russell & Bullock, 1985). The goodness-of-fit of the MDS representation was estimated with the Stress measure. We expected that faces clustered according to their similarity in emotional expression rather than identity in the bidimensional face space.

Visual similarity. We measured the visual similarity among faces by computing the Pearson correlations between pixel values of light intensity for each pair of faces. This was done to exclude the possibility that differences in similarity judgements among conditions were due to visual similarity. In particular, we first prepared the dataset of images by transforming them into grey scale and applying histogram equalization to enhance the contrast of the image and maximize the prominence of discernible features. Second, each image was converted into an n*n matrix, where n is the number of pixels of the image and each entry represented a pixel value of light intensity. Then, we computed a covariance matrix (n*n)*m of the set of images (where m is the total number of images), and transformed it into correlational matrix. In order to obtain always positive values, we converted the correlation coefficients into correlational distances (1 - Pearson correlation). These were entered in a 20 x 20 representational dissimilarity matrix, wherein the row and the columns represented the faces and each cell the correlational distance between faces in each specific pair. We extracted from this matrix the mean and the standard deviation of each condition of interest, which resembled those in the similarity ratings matrix. These were used as dependent variables in a one-way repeated- measures ANOVA, wherein we used a planned contrast to test the same main hypothesis, that is, lower similarity (higher dissimilarity) in ID compared to EE, NN and EE_DF ($p<0.05$).

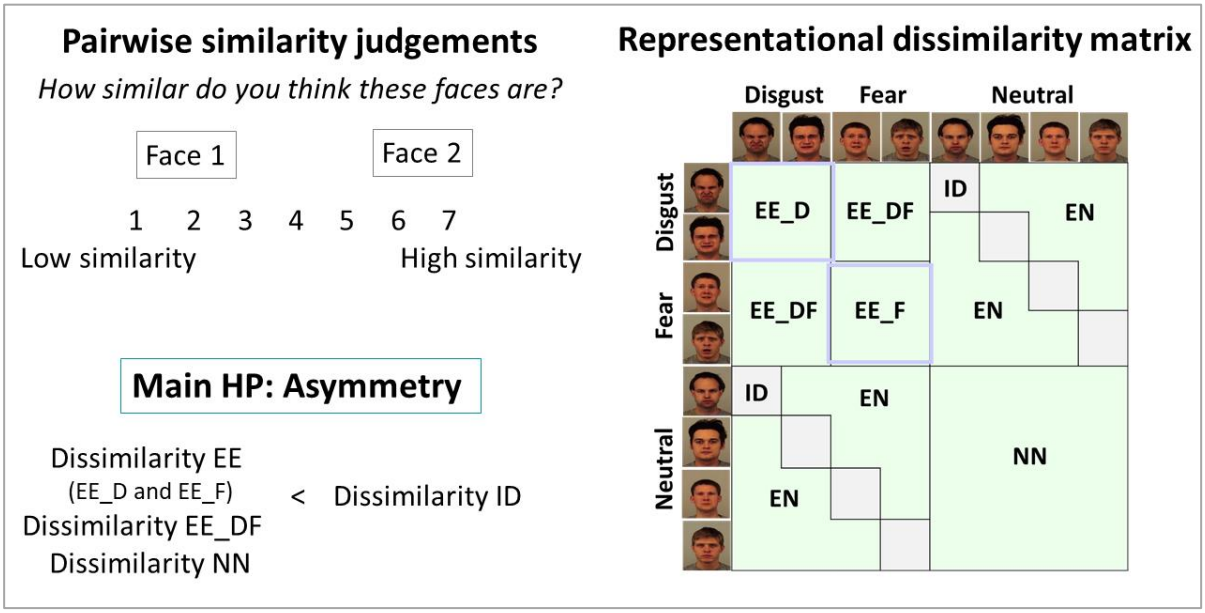


Figure 3.1. Graphical representation of the task structures, conditions of interest and key hypotheses. During the task, participants rated the similarity among all the possible combinations of 20 faces (5 disgust, 5 fear, 10 neutral) on a 7 points scale (1=low similarity, 7 high similarity). The similarity ratings were standardized, transformed into dissimilarity measures (correlational distance) and entered in a 20 x 20 representational dissimilarity matrix (RDM). In the RDM, the rows and the columns represented the stimuli (disgust: 1 to 5; fear: 6 to 10; neutral: 11 to 20), and each cell a correlational distance between faces in each pair. In the RDM, the violet squares represent the dissimilarity within emotional pictures (EE), calculated by averaging the dissimilarity within disgusted (EE_D) and fearful (EE_F) faces; EE_DF, is the dissimilarity between disgusted and fearful faces, and NN, the dissimilarity within neutral faces; ID, depicted in grey colour, indicates the dissimilarity between emotional and neutral faces, with the same identity, and EN the dissimilarity between emotional and neutral faces, with different identities. We expected an asymmetric effect of emotional expression and identity on similarity judgements, resulting in higher similarity (lower dissimilarity) in EE, EE_DF and NN compared to ID.

3.3 Results

In contrast to our hypothesis, we did not observe lower similarity ratings in ID compared to EE, $F(1, 19)=1.80$, $p=0.20$, $\eta_p^2=0.09$; EE_DF, $F(1, 19)=0.65$, $p=0.43$, $\eta_p^2=0.03$; and NN, $F(1, 19)=2.30$, $p=0.15$, $\eta_p^2=0.11$), indicating a symmetric rather than an asymmetric effect of emotional expression and identity on similarity judgements. Crucially, in post-hoc tests we did observe higher dissimilarity in EE_DF than NN ($p<0.001$) and EE ($p<0.001$), suggesting that any mismatch in emotion results in lower similarity judgements. As predicted, the manipulation check revealed higher dissimilarity in EN compared to EE, $F(1, 19)=77.83$, $p<0.001$, $\eta_p^2=0.80$; EE_DF, $F(1, 19)=35.34$, $p<0.001$, $\eta_p^2=0.65$; NN, $F(1, 19)=54.49$, $p<0.001$, $\eta_p^2=0.74$; and ID, $F(1, 19)=37.20$, $p<0.001$, $\eta_p^2=0.66$. Given the small sample size, we calculated the inter-raters reliability, which resulted in a very good Cronbach's Alpha ($\alpha= 0.97$). We also measured the visual similarity

among faces by computing the correlational distance among pixel values of light intensity for each pair of faces, in order to exclude the possibility that differences in similarity judgements among conditions were due to visual similarity. We found higher visual similarity (lower correlational distance) in ID compared to EN_DF, $F(1, 9)=33.93$, $p<0.001$, $\eta_p^2=0.79$; and NN, $F(1, 9)=18.02$, $p=0.002$, $\eta_p^2=0.67$, but only a trend towards significance was observed between ID and EE, $F(1, 9)=4.01$, $p=0.08$, $\eta_p^2=0.31$. The MDS solution showed that the faces were clustered according to their similarity in valence and arousal (but not visual similarity) in a two-dimensional space. It had a Stress value of 0.04, indicating a good fit for this model. These findings are reported in Figure 3.2.

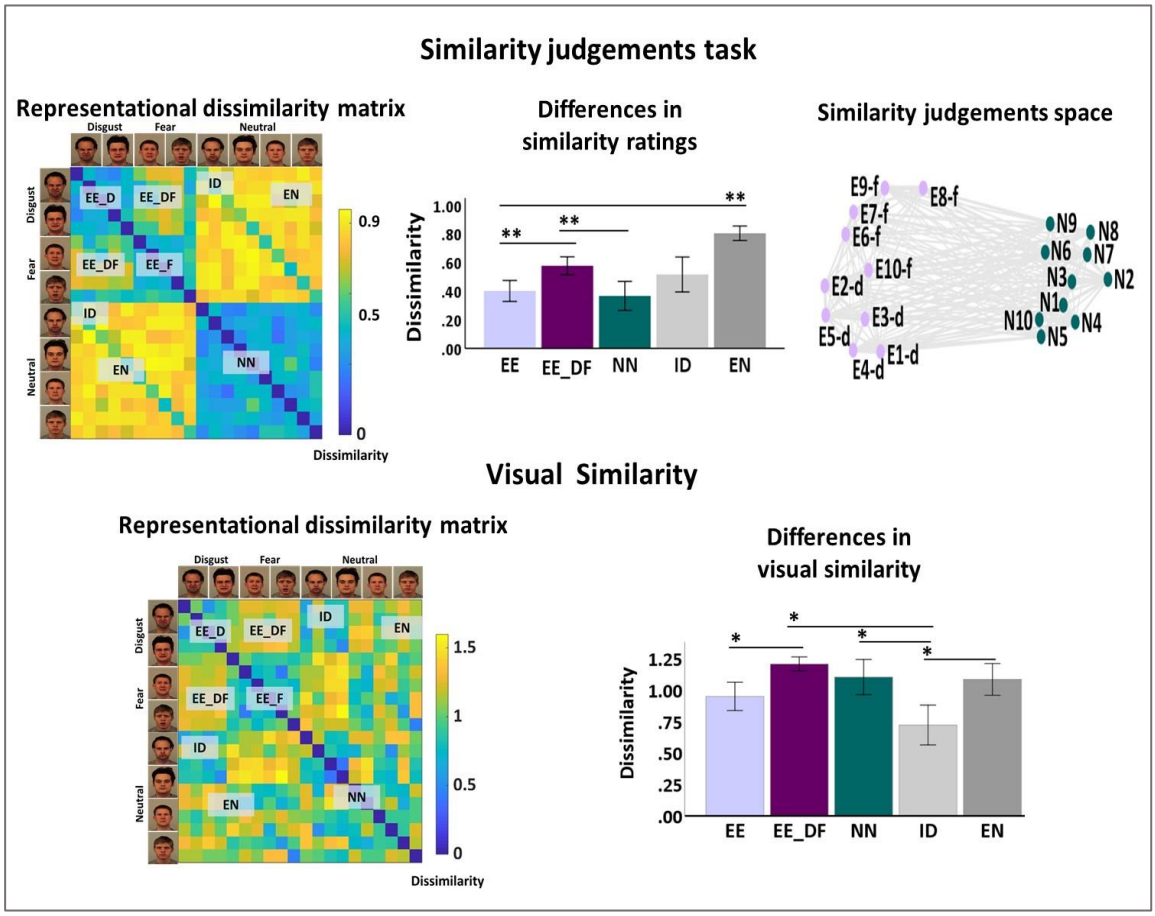


Figure 3.2. Top. Similarity Judgements task. Left: Representational Dissimilarity Matrix (RDM) of the similarity ratings of 20 faces (5 disgust, 5 fear, 10 neutral), averaged across participants. It is symmetric about a diagonal of zeros, the rows and the columns represent the stimuli, and each cell the dissimilarity (measured as correlational distance: 1- standardized similarity ratings) between stimuli within each specific pair. Yellow colours denote high dissimilarity, blue colours low dissimilarity. Centre: differences in dissimilarity (measured as correlational distance) among conditions during the similarity judgements task: average dissimilarity within emotional faces (averaged across disgusted and fearful faces) (EE), between emotional faces (EE_DF), within neutral faces (NN), between emotional and neutral faces of the same identity (ID), and between emotional and neutral faces (with different identities) (EN). Right: The Multidimensional Scaling (MDS) plot of the 20 faces in a bidimensional space. Stimuli from E1_d to E5_d represented 5 disgusted faces, from E6_f to E10_f 5 fearful faces, and from N1 to N10 10 neutral faces. Bottom. Visual similarity. Left: Representational Dissimilarity Matrix (RDM) of the visual similarity of 20 faces (5 disgust, 5 fear, 10 neutral). It is symmetric about a diagonal of zeros, the rows and the columns represent the stimuli, and each cell the correlational distance between stimuli within each specific pair. Yellow colours denote high dissimilarity, blue colours low dissimilarity. Right: differences in visual similarity (measured as correlational distance) among conditions. Error bars represent ±2 SEM; *p_{FWE}<0.05; **p_{FWE}<0.001. Abbreviations: E, Emotional; f, fearful faces; d, disgusted faces; N, neutral faces.

3.4 Discussion

In this study, we investigated the asymmetric effect of emotional expression and identity on the perception of similarity between faces. We explored whether participants relied more on emotional or visual features while judging the similarity between emotional and neutral faces, without instructing them on which aspect to focus. We report two new findings. First, emotional and visual features had the same relevance in similarity judgements: Mr Hyde is perceived as equally similar to Jack the Ripper and to his alter ego Dr Jekyll. This result suggests a symmetric rather than an asymmetric effect on similarity perception. Second, similarity ratings were not fully explained by the visual similarity of faces, evident in that NN and EE were less visually similar (higher correlational distance) than ID, yet participants did not perceive these conditions to be different from each other in similarity. We also found that emotional similarity among faces may influence overall similarity perception, given the higher dissimilarity in conditions with an emotional mismatch (i.e., EE_DF and EN) compared to those with emotional congruency (i.e., EE and NN). Below we discuss the implications of these findings.

Symmetrical effects of emotional and visual features on similarity judgements provide additional evidence for the relevance of emotion in similarity judgements. Further support comes from the observation that an emotional mismatch (i.e., EN and EE_DF conditions) makes people perceive faces as less similar compared to conditions with emotional congruency (i.e., EE and NN). As previously proposed (Jamin Brett Halberstadt & Niedenthal, 1997; Wegrzyn et al., 2017), this process is evolutionarily advantageous: poor discrimination among emotional expressions that have the same meaning (expressions of disgust, for example) possibly would not endanger the individual; however, when the stimulus is emotional, small dissimilarities can create large differences in similarity perception and action planning (e.g., fight or flight). Disgusted and fearful faces in the EE_DF condition have similar values in valence and arousal (low scores in valence and high scores in arousal). This is also the case for the neutral faces in the NN condition (medium scores in both valence and arousal). Yet, small variations in valence and arousal were more relevant when the faces were emotional, rather than neutral. Emotions convey specific information about one's internal and external environment that each individual takes into consideration for congruent action planning and decision

making. This is made possible by selectively focusing attention to the emotional aspects of the world, and it will probably result in a lower latency in detecting the emotional content of any stimulus and increased discriminability of stimuli exhibiting those features. Furthermore, for the first time we disentangled emotional and visual similarity, by computing the latter in an objective rather than in a subjective way. We observed lower visual similarity between emotionally similar faces, both neutral and negative emotional, than between faces with the same identity but different emotions. This suggests that the symmetric effect observed in the similarity ratings task is not explained by visual similarity.

Our study has several limitations that can be addressed in future work. First, we studied only two negative emotions, neglecting positively-valenced emotions. We chose this to increase the statistical power in terms of number of trials per condition, while keeping the experiment short enough to ensure participants' attention. It would be relevant in future studies to examine whether the same effects are replicated with positive emotions. Second, we only selected male facial expressions. This was a deliberate choice, to ensure that participants would focus on the visual and emotional similarity among faces. However, it would be interesting to include gender as an additional dimension in face space and to explore its relative weight in similarity judgements. Finally, our sample size was quite small, even though the inter-raters reliability was very high. However, given the significance and applicability of the findings, it would be appropriate to replicate the experiment by increasing its sample size, and by including an equal number of male and female faces. This would test whether gender moderates the previously reported effects.

Overall, in the present study we report a symmetrical effect of emotional expression and identity on similarity judgements. Mr Hyde is equally similar to Dr Jekyll and to Jack the Ripper, despite the higher visual similarity to the latter. Determining the relative importance of identity and emotion in similarity judgements is of paramount importance to combine the emotion cognition and computer science literatures to fill the 'affective gap' between facial visual features and their psychological representations in human observers (Zhao et al., 2018).

4. Chapter: The neural representations of emotional experiences are more similar than those of neutral experiences

This chapter has been accepted for publication in the *Journal of Neuroscience* (2022).

Abstract

Stimuli that evoke the same feelings can nevertheless look different and have different semantic meanings. Although we know much about the neural representation of emotion, the neural underpinnings of emotional similarity are unknown. One possibility is that the same brain regions represent similarity between emotional and neutral stimuli, perhaps with different strengths. Alternatively, emotional similarity could be coded in separate regions, possibly those sensitive to emotional valence and arousal. In behaviour, the extent to which people consider similarity along emotional dimensions when they evaluate the overall similarity between stimuli has never been investigated. While the emotional features of stimuli may dominate explicit ratings of similarity, it is also possible that people neglect emotional dimensions as irrelevant to that judgement. We contrasted these hypotheses in (male and female) healthy controls using two measures of similarity and two picture databases of complex negative and neutral scenes, the second of which afforded exquisite control over semantic and visual attributes. The similarity between emotional stimuli was greater than between neutral stimuli in the inferior temporal cortex, the fusiform face area, and the precuneus. Additionally, only the similarity between emotional stimuli was significantly represented in early visual cortex, anterior insula and dorsal anterior cingulate cortex. Intriguingly, despite the stronger neural similarity between emotional stimuli, the same participants did not rate them as more similar to each other than neutral stimuli. These results contribute to our understanding of how emotion is represented within a general conceptual workspace and of the overgeneralisation bias in anxiety disorders.

4.1 Introduction

We may judge an image of a homeless person and of a car accident as different because of their different meanings, or as similar because both evoke negative feelings. Emotional similarity refers to the tendency to group stimuli together because they evoke the same feelings (Martina Riberto et al., 2019). The extent to which similarity along emotional dimensions influences perceived similarity between complex experiences is unknown. It is important to understand the effect of emotion on similarity because aberrant similarity perception influences psychological well-being (Puccetti et al., 2021) and is clinically relevant in anxiety and posttraumatic stress disorders (Laufer et al., 2016). For example, after a traumatic event patients may consider later experiences to be similar to the original, fearful one not because of their ostensible meaning, but their emotional similarity.

All stimuli can be described according to their location on orthogonal dimensions, valence and arousal, with their proximities reflecting aspects of their relationship (Russell, 1980). This perspective suggests that entirely-neutral stimuli, at the axes' origin, may be perceived to be just as similar as stimuli at the extremes. Yet similarity inferred from single-stimulus judgements on single attributes (e.g., shape, valence) rarely explains more than half the variance in explicit ratings of similarity (Iordan, Ellis, Osherson, & Cohen, 2017). Indeed, highly-arousing negative stimuli may be perceived less similar to each other than neutral ones if they evoke qualitatively different emotions (e.g., fear, anger). Previous comparisons reveal increased ratings of similarity among negative pictures than among randomly-selected neutral pictures (Talmi, 2013) and among positive than negative stimuli (Koch et al., 2016). Unfortunately, previous rating studies employed semantically-related emotional stimuli, thereby confounding emotion and semantic similarity. Nevertheless, in conditioning paradigms, where semantic similarity is not a confound, wider generalisation of aversively-conditioned stimuli has been observed (Laufer & Paz, 2012). Therefore, we hypothesized that negative emotional stimuli will be perceived as more similar than neutral stimuli.

Recent neuroimaging studies observed low specificity for discrete emotions, and provided evidence against a locationist perspective to the study of emotions (Hoemann et al., 2019). Instead, emotional stimuli are likely represented in

distributed networks of cortical and subcortical regions, which are not functionally specific to affect (Chang, Gianaros, Manuck, Krishnan, & Wager, 2015) but carry out emotion-relevant computations: the occipitotemporal regions, visual-semantic processing of emotional and neutral categories (Kragel, Reddan, LaBar, & Wager, 2019); the insula and the anterior cingulate cortex, awareness of bodily sensations and visceral regulation necessary for a core affective state representation; and the ventral prefrontal cortex, positive valence (Lindquist et al., 2012). No previous work has directly compared the neural underpinnings of emotional and neutral similarity for complex, realistic stimuli, but a handful of studies employed simple stimuli. Representational similarity analysis (RSA) maps similarity perception in the brain by correlating neural and behavioural data (Kriegeskorte, Mur, et al., 2008b). This technique revealed increased neural similarity between conditioned stimuli in the amygdala (R. M. Visser, Scholte, Beemsterboer, & Kindt, 2013), the occipitotemporal cortex (Dunsmoor et al., 2013) and the superior frontal gyrus (R. M. Visser, Scholte, & Kindt, 2011), and increased similarity between stimuli that predict reward (Zeithamova, Gelman, Frank, & Preston, 2018) and pain (Wagner, Rütgen, & Lamm, 2020) in the hippocampus. Following from this theoretical and empirical work, we hypothesized that neural similarity will differ as a function of stimulus emotionality. Specifically, we hypothesized that the neural similarity among emotional stimuli will be greater than among neutral stimuli - expressing the predicted pattern of behavioural ratings. Emotion may increase neural similarity in any region that encodes participants' self-reported similarity space, but may do so more strongly in regions that serve emotion-relevant operations.

We tested these hypotheses in a series of experiments that present several strengths compared to the state-of-the-art. We used different similarity judgements tasks and picture databases, one of which permitted, for the first time, control over taxonomic and thematic similarity, and narrowed our search volume through innovative searchlight approaches.

4.2 Materials and Methods

Participants

A total of 90 participants were recruited from the University of Manchester (UK), and from the Weizmann Institute of Science (Israel) to take part in the study (age range, 20–54 years; mean age, 30.14 years; SD, 7.17) (Experiment 1: 20 participants, 10 females; Experiment 2: 40 participants, 20 females; Experiment 3: 29 participants, 12 females; one participant was excluded, because he did not follow the instructions of the task). The sample size was selected according to previous publications in this research field (Charest et al., 2014; Chikazoe et al., 2014; Giordano et al., 2021). All participants had normal or corrected-to-normal vision, and were older than 18 years. They gave informed consent prior to the experiment and have been reimbursed for their participation (£5 for the behavioural experiments, £22 for the MRI experiment). The exclusion criteria were: a history of neurological (e.g., head injury or concussion) or psychiatric conditions (e.g., depression, anxiety), drug or alcohol abuse, or regular medication that could influence emotional processing. The study was approved by the ethics board of the University of Manchester and of the Weizmann Institute of Science (protocol number 0287–09-TLV).

Materials

First database of complex pictures

In experiment 1, we selected 20 images taken from the Nencki Affective Picture System (NAPS) database (Marchewka et al., 2014). Picture IDs that we selected in experiment 1 are reported in Figure 4.2-1. NAPS has been validated for use in emotional research (Riegel et al., 2016; Wierzba et al., 2015) and consists of 1,356 realistic, high-quality photographs divided into five categories (people, faces, animals, objects, and landscapes). In order to control for visual similarity, we matched the pictures for low-level visual features, that, unlike subjective ratings of visual complexity, are not affected by the arousal complexity bias (Madan et al., 2018) and by the vividness bias (Todd, Schmitz, Susskind, & Anderson, 2013). These measures included the luminance (the average pixel value of the greyscale image) and the contrast (the standard deviation across all the pixels of the greyscale image) (Bex & Makous, 2002). In order to quantify the colours within each image, we computed the quantity of red (R), green (G) and blue (B), according to the RGB colour model. Finally, the JPEG size and the entropy of each

greyscale image were used as indices of the overall visual complexity of each image (Donderi, 2006). The JPEG size was determined with a compression quality setting of 80 (on a scale from 1 to 100). Perceptually simple images are highly compressible and therefore result in smaller file sizes. The entropy, H, is computed from the histogram distribution of the 8-bit grey-level intensity values x: $H = -\Sigma p(x) \log p(x)$, where p represents the probability of an intensity value x. H varies with the 'randomness' of an image. High-entropy images are noisier and have a high degree of contrast from one pixel to the next, whereas low-entropy images have rather large uniform areas with limited contrast. The sample of images included 10 emotional and 10 neutral images. The designation of images to this category was based on the NAPS ratings of valence and arousal on a 9-points scale provided by 204 European participants. We considered emotional pictures as rated less than 4 in the valence scale (negative valence) and more than 6 in the arousal scale (high arousal), whereas the neutral images ranged from 4 to 6 in both dimensions. To validate the NAPS norms, we also asked our participants to rate the valence and the arousal of the picture before the main task. Figure 4.2-1 (supplementary information, SI) showed the picture IDs from the NAPS database ('people' category), divided into emotional and neutrals. Table 4.1 showed the mean and the standard deviation of the different visual and emotional measures for emotional and neutral pictures, as well as the differences between them. We controlled to some extent for semantic similarity, namely the similarity both in the theme (e.g., violence) each picture depicts, other categories it belong to (e.g., outdoor scene), and its specific meaning. With this aim, we choose images that included more than one person in an outdoor scene from the same category - the 'people' category. These images contain a lot of information beyond the persons themselves, placing them in a rich and realistic context. The matching we achieved between emotional and neutral conditions exceeds that in most published studies and represents the current state-of-the-art in controlling emotional and neutral stimuli in research. However, the range of emotional themes was reduced compared to that in the neutral set. Therefore, emotional pictures might be rated as more similar, because of the higher thematic similarity compared to neutrals.

		Categories		Statistics		
		Emotional	**Neutral**	**t**	**p**	**d**
Visual measures	**Luminance**	89.42 ± 27.31	110.13 ± 36.06	-1.45	.16	-3.68
	Contrast	61.97 ± 1.16	62.19 ± 1.14	-.49	.96	-0.07
	R	93.24 ± 26.06	111.62 ± 26.67	-1.48	.17	-3.48
	G	88.99 ± 27.92	110.12 ± 37.84	-1.41	.19	-3.69
	B	82.58 ± 26.54	105.95 ± 41.24	-1.56	.15	-4.04
	JPEG	337121.90 ± 90579.45	277070.50 ± 69785.15	1.66	.11	3.86
	Entropy	7.50 ± .28	7.49 ± .24	.11	.91	0.02
Emotional measures	**Valence**	1.985 ± .743	4.840 ± .976	-10.22	.000**	-3.09
	Arousal	7.170 ± 1.389	4.840 ± 1.219	8.93	.000**	2.04

Table 4.1. Differences in visual and emotional measures between emotional (n=10) and neutral (n=10) pictures (experiment 1). The mean and the standard deviation of each measure are shown, as well as the t, the p value and Cohen's d as effect size measure for each difference. **, pFWE< 0.001.

Second database of complex pictures

In experiments 2-3, in order to control the emotional and neutral pictures for thematic similarity, we selected natural scenes in a way that all the categories depicted realistic events that do not co-occur in the environment. In particular, we chose 72 real-world colour photographs using Google images, which represented one or more people in outdoor situations. We divided them into 4 categories according to the scene that was depicted, resulting in 18 images per category. Two of the categories were neutral, and two were emotionally-arousing and negatively valenced. These latter categories represented either poverty scenes (emotional category 1, E1) or car accidents (emotional category 2, E2). The neutral categories portrayed either people hanging laundry to dry (neutral category 1, N1) or talking on the phone (neutral category 2, N2). The full set of pictures can be found in Figure 4.3-1 (SI). We minimised the thematic similarity between emotional categories, by selecting for each of the emotional categories action-context combinations that do not normally occur in a common theme or scenario. The same was true for the two neutral categories. To control for taxonomic similarity to some extent, all the pictures we selected shared two semantic features, they depicted *people outdoor*. Second, we controlled the pictures for affordance, namely the action that a scene can afford, by selecting pictures that depicted only one type of action – and therefore, affordances - in each category. Specifically, in

E1, people sit on the ground while begging; in E2, accident victim(s) lay either on a surface (the ground) or a crashed car; in N1, people stand hanging and drying clothes; in N2, they stand or walk talking on the phone in the street. Although these actions and affordances differed across the 4 categories, the design ensures that these differences did not influence comparisons across the two neutral and two emotional categories. Finally, we controlled the stimuli for visual properties, as in experiment 1. An independent sample of 10 healthy participants rated the valence and the arousal of the stimuli, and another independent sample of 20 participants judged the similarity of the pictures. Table 4.2 showed the mean and standard deviation of visual and emotional measures for each category as well as the differences among them. Table 4.3 showed the mean and standard deviation of similarity measures within and between categories, as well as the differences among them.

	Categories				**Statistics**		
	E1	**E2**	**N1**	**N2**	**F**	**p**	**η_p^2**
Luminance	105.39 ± 21.37	95.68 ± 26.27	106.45 ± 29.79	106.22 ± 25.11	0.73	0.54	0.31
Contrast	58.63 ± 10.80	62.52 ± 6.09	64.99 ± 11.12	63.54 ± 12.69	1.22	0.31	0.05
R	115.07 ± 20.73	98.84 ± 26.71	114.15 ± 31.11	109.28 ± 26.30	1.47	0.23	0.06
G	102.18 ± 21.99	94.14 ± 26.23	104.88 ± 30.71	105.40 ± 26.07	0.70	0.55	0.03
B	96.50 ± 23.29	95.35 ± 28.25	94.32 ± 28.11	102.96 ± 28.94	0.37	0.77	0.02
Jpeg	66701.89 ± 11078.33	63614.83 ± 11967.82	59643.05 ± 12220.83	67011.28 ± 28005.71	0.70	0.55	0.03
Entropy	7.58 ± 0.25	7.56 ± 0.26	7.54 ± 0.29	7.52 ± 0.22	0.20	0.90	0.00
Valence	2.91 ± 1.42	1.97 ± 1.02	4.91 ± .26	5.13 ± .30	46.93	0.00	0.84
Arousal	6.64 ± 1.40	7.74 ± 1.34	4.72 ± 1.37	4.53 ± 1.26	27.37	0.00	0.75

Table 4.2. Differences in visual and emotional measures among categories. The mean and the standard deviation of each measure are shown, as well as the F, the p value and the partial eta squared for each difference. Abbreviations: E1, Emotional category 1 (poverty scenes, n=18); E2, Emotional category 2 (car accidents, n=18); N1, neutral category 1 (laundry scenes, n=18); N2, neutral category 2 (telephone scenes, n=18) (experiment 2-3).

Dissimilarity Within category				Dissimilarity Between categories		
E1	**E2**	**N1**	**N2**	**E12**	**N12**	**EN**
.004±.002	.003±.003	.004±.003	.004±.003	.017±.003	.019±.002	.024± .001

Statistics Manipulation check			Statistics Main HP		
F	**p**	**η_p^2**	**F**	**p**	**η_p^2**
292.56	<0.001	0.94	4.11	0.06	0.18

Table 4.3. Differences in dissimilarity among categories (validation study). The mean and the standard deviation of each condition of interest are shown, as well as the F, the p value and the partial eta squared for each difference. Abbreviations. Dissimilarity within: E1, emotional category 1 (poverty scenes, n=18); E2, Emotional category 2 (car accidents, n=18); N1, neutral category 1 (laundry scenes, n=18); N2, neutral category 2 (telephone scenes, n=18). Dissimilarity between: E12, emotional categories; N12, neutral categories; EN, emotional and neutral categories.

Experimental design

A graphical representation of the general experimental design is shown in Figure 4.1. In all the experiments, we asked participants to judge the similarity of a set of complex pictures to test our main hypothesis for the behavioural data: that the perceived similarity between emotional compared to between neutral pictures will be higher. As shown at the top of Figure 4.1, in the first two experiments participants performed a pairwise similarity rating task. In experiment 1, after rating the valence and arousal of each picture from the first dataset, participants rated all the possible combinations among the stimuli. In experiment 2 we focused on the ratings of interest (denoted with red circles at the bottom of Figure 4.1), and therefore, participants only rated the similarity between emotional categories (E12) and between neutral categories (N12) of pictures from the second dataset, as well as between emotional and neutral categories (EN), with the latter pairs serving as catch trials. Experiments 1-2 ended after approximately twenty minutes. In experiment 3, after a functional Magnetic Resonance Imaging (fMRI) scan, participants performed a surprise multi-arrangements (MA) task to judge the similarity of the 72 pictures on a bidimensional space, as depicted at the top right of Figure 4.1 (duration: approximately one hour).

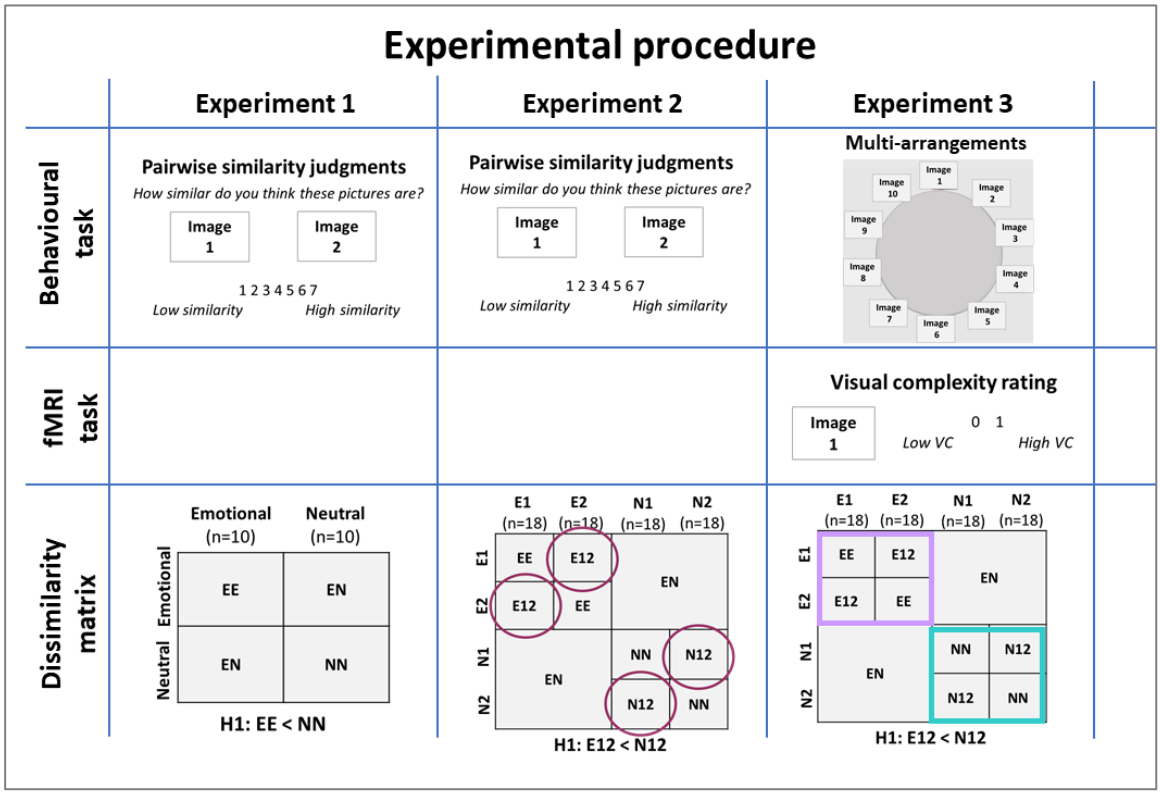


Figure 4.1. Graphical representation of the experimental procedure. In experiments 1-2, participants performed the same behavioural task. They were presented with a pair of pictures and rated their similarity on a 7-points scale (low to high similarity). In experiment 1, participants judged all the possible combinations from the 1st database, which consisted of 20 complex pictures (10 emotional and 10 neutral) selected from the NAPS. We expected as main finding lower dissimilarity (higher similarity) between emotional (EE) than neutral (NN) pictures. In experiment 2, participants judged the similarity between emotional and neutral pictures from the 2nd database. It consisted of 72 pictures from 4 semantic categories (18 pictures in each category), two emotional (E1 and E2) and two neutral (N1 and N2). Participants only rated E12, N12 and few EN pairs only: E12 represented the similarity between E1 and E2, N12 between N1 and N2, and EN between emotional and neutral pictures. We expected lower dissimilarity (higher similarity) in the former. In both experiments 1-2, EN comparisons served as manipulation checks. The same database was used in experiment 3, wherein participants first judged the subjective visual complexity of each picture during a functional magnetic resonance imaging (fMRI) scan, and then judged the similarity among all the pictures by arranging them in a circular arena. We tested the same hypothesis as in experiment 2, and extended it also to the neural data. The violet square in the dissimilarity matrix represents the 'emotional similarity space', and the green one the 'neutral similarity space'.

Valence and arousal rating task

The two dimensions of valence and arousal are considered to be key to the conceptual representation of semantic concepts as well as emotional stimuli. Therefore, we used these for stimulus selection, such that those selected for the emotional condition differed from those selected to the neutral condition along both valence and arousal dimensions. To validate the designation of pictures from the two datasets to emotional and neutral conditions, participants completed a valence and arousal rating task, following the procedure suggested by Lang et al. (2008) (Lang et al., 2008). Each trial started with a central fixation cross for 500 ms. Then,

participants viewed one of images presented in the centre of the screen, and rated each pictures on two 9-points scale (valence scale: 1, negative emotions; 9, positive emotions; 5 neutrals. Arousal scale: 1, relaxed; 9, aroused; 5 neutral). We instructed participants to respond as quickly as possible by clicking the appropriate number key, and informed them that there was not a right or wrong answer. Pictures from the first dataset were rated by participants in experiment 1 prior to commencing that experiment, while the ratings of pictures from the second dataset was completed by a separate group of participants.

Behavioural measures of similarity

The data from the behavioural experiments were used as measures of perceived similarity, that is, similarity ratings in experiment 1-2, and Euclidean distance in experiment 3. To make sure that the behavioural findings were independent of the specific instructions participants were given two separate task instructions (pairwise ratings in experiment 2, multi-arrangement in experiment 3).

Pairwise similarity rating task

In experiment 1-2, participants rated the similarity of paired pictures on a 7-points scale (1= low similarity, 7= high similarity). In experiment 1, they rated all possible pairwise combinations (190 pairs), resulting from the database of 20 complex pictures. In experiment 2, because of time constraint, we divided the 72 pictures into two subsets ('even' and 'odd', n=36 within each subsets); in addition, we focused on pairs in E12 and N12, as well as some in EN as catch trials (total pairs= 170; 81 in both E12 and N12, and 8 in EN). We chose the pairwise presentation, because each pair is independently rated and also small differences in similarity judgements can be detected, compared to a triad 'forced-choice' similarity task, wherein only binary responses are provided (Goldstone et al., 1997; Miller, 1994). We instructed participants to base their judgment on the overall meaning of the picture, without considering any visual details (e.g., the background colour, the number of people). We also informed them that there was not a right or wrong answer. We purposefully did not bias them to emphasize any dimension because we wanted our laboratory measure of behavioural similarity perception to quantify, as closely as possible, 'natural', holistic similarity perception outside the lab.

Finally, to make sure that the behavioural findings were independent of the specific instructions, participants followed two separate task instructions (pairwise ratings in experiment 2, multi-arrangement in experiment 3).

Multi-arrangements task

In the validation study of the second database and in experiment 3, participants judged the similarity among all the pictures by using the multi-arrangements (MA) task. We chose it because it is a quick and efficient task for acquiring similarity judgements in experiments with a relatively large number of stimuli. Kriegeskorte and Mur (2012) established the MA test-retest reliability (r=0.81) as well as external validity (Kriegeskorte & Mur, 2012). The task comprised different trials. In each trial, a subset of 16 stimuli was presented along the perimeter of a circle, or 'arena', on a computer screen. Participants had unlimited time to drag and drop the stimuli in the arena according to their similarity, such that similar stimuli were placed close to each other and dissimilar stimuli apart. In other words, the distance among stimuli in the arena reflected their dissimilarity. We instructed participants to focus on the content of the pictures and to ignore visual details (e.g., the colour of the background, the number of people in the scene). A trial ended when participants arranged all the stimuli in the arena. Subsequent trials started with another subset of stimuli to be arranged, selected by using the 'Lift-the-weakest algorithm for adaptive design of item subsets' (Kriegeskorte & Mur, 2012). This method optimises trial efficiency by adaptively selecting item subsets whose dissimilarity estimates presented the weakest evidence. The task ended after approximately one hour, when participants judged all the possible combinations among stimuli.

MRI procedure

In experiment 3, images were acquired on whole body MRI scanner (Trio TIM, Siemens, Germany) with a 12-channel head coil. Functional images were acquired with a susceptibility weighted EPI sequence (TR/TE=2000/30 ms, flip angle=75 degrees, voxel dimensions=3x3x3.5 mm, 192 slices) in 4 separate scanning sessions (up to two minutes between sessions). Anatomical T1-weighted images were acquired after the functional scans (MPRAGE, Repetition time (TR)/Inversion

delay time (TI)/Echo time (TE)=2500/900/2.32 ms, flip angle=8 degrees, voxel dimensions=1 mm isotropic, 32 slices).

As shown in Figure 4.1, during the fMRI scan, participants viewed the 72 complex pictures on a blank screen (size 800 x 800 pixels, visual angle: 64°); we asked them to rate their visual complexity to make them focus on the stimuli, by pressing the right or the left button of the response box, respectively. Images were presented in a random order for 3 seconds, during which participants had to make their ratings, interleaved with a black fixation cross (mean jitter 3 seconds). The task was divided into 4 runs, during which every picture was presented once, thus resulting in 4 repetitions for each picture, and a total duration of approximately 50 minutes. We instructed participants that there was not a right or wrong answer in the task; rather, they had to focus on their *subjective* perception during the ratings. In order to guide participants in the ratings, we suggested to them that 'a picture of few objects, colours, or structures would be less complex than a very colourful picture of many objects that is composed of several components' according to Madan et al. (2018). Behavioural and fMRI tasks instructions differed, as it is not possible to measure both neural representational similarity and behavioural similarity using the same instructions. Similar procedures were also adopted in previously published papers in this research field (Chavez & Heatherton, 2015; Chikazoe et al., 2014; Kriegeskorte, Mur, Ruff, et al., 2008). This is because to compute the neural representation of each picture (and then feed it into the RSA), in the MRI session we need participants to focus on one picture at a time; but behavioural measures of similarity perception requires that participants consider picture pairs.

Statistical data analysis

In the similarity judgements tasks, we expected higher similarity (lower dissimilarity) within-category than between categories. We also expected higher similarity (lower dissimilarity) between emotional than between neutral conditions, as showed at the bottom of Figure 4.1. The first prediction serves as manipulation check, since a good category boundary simultaneously maximize the within-category similarity, and minimize the between categories similarity; the second prediction represents our main hypothesis, and applies also for the neural data. In experiment 1, EN was calculated by averaging the dissimilarity between emotional

and neutral pictures, and the dissimilarity within-emotional (EE) and within-neutral (NN) categories by averaging the dissimilarity between emotional, and between neutral pictures, respectively, for each participant. In experiment 3, EE represented the averaged dissimilarity within E1 and within E2, NN the averaged dissimilarity within N1 and within N2, and EN across both E1 and E2, and N1 and N2, for each participant. Finally, in experiments 2-3, E12 was measured by averaging the dissimilarity between the two emotional categories, and N12 between the two neutral categories. The conditions of experiment 3 are the same in the validation study. Additional details about the statistical analyses are reported in the following sections.

Behavioural data analysis

We analysed these data by using Representational similarity analysis (RSA). Specifically, in experiment 1, the similarity ratings were entered as input in a 20 x 20 similarity matrix for each participant. The rows and the columns represented the experimental stimuli, and each cell reflected the similarity rating for each pair. Then, for each subject, a Representational Dissimilarity Matrix (RDM) was computed. We first standardized the similarity ratings, by subtracting 1 (the lowest similarity rating) from each rating *x*, and then divided by 6 (highest similarity rating - lowest similarity rating). Second, we transformed them into correlational distances, by subtracting the ratings from 1. The correlational distance ranges from 0 to 2 (0 for perfect correlation, and thus high similarity; 1 for no correlation; 2 for perfect anticorrelation), and was entered as input in each cell of the RDM. As a consequence, the RDM is symmetric about a diagonal of zeros. Next, we extracted from the single-subject RDM the mean dissimilarity and the standard deviation of the conditions of interest as mentioned in the key hypotheses. These were entered as dependent variables in a repeated-measures ANOVA, with the conditions as grouping factor (experiment 1: EE, NN, and EN; experiment 2: E12, N12, EN). In the validation study and in experiment 3 (MA task), similarity was measured as Euclidean distance between stimuli in the arena. Specifically, at the end of each trial, a *partial* RDM is estimated, showing the Euclidean distance between stimuli within each trial. At the end of the task, a *global* 72 x 72 RDM is estimated by averaging the partial RDMs with an iterative rescaling. This scaling procedure takes into account that in each trial participants focused on a specific

subset, and that, therefore, there is not a permanent relationship between screen distance and dissimilarities across trials (see (Kriegeskorte & Mur, 2012) for details). Then, we extracted from each participant's *globa*l RDM the mean and the standard deviation of the conditions of interest mentioned in the section about the key hypotheses. These were entered as dependent variables in a repeated-measures ANOVA, which served to test lower dissimilarity in EE and NN than in E12, N12 and EN, and the main hypothesis (lower dissimilarity in E12 than N12). Bonferroni post hoc corrections for multiple comparisons ($p<0.05$) were used to explore the nature of the effect. The results of the validation study are shown in Table 4.3.

We conducted additional analyses to test differences in the variance across participants in the judgements of similarity between emotional than neutral stimuli. With this aim, we conducted two-samples F-tests for variance, one for each contrast of interest: experiment 1: EE vs NN; experiment 2: E12 vs N12; experiment 3: EE vs NN, and E12 vs N12.

Multidimensional scaling (MDS). We performed the MDS to visualise the structure of the similarity space, wherein proximities reflect similarities among stimuli and are measured on an ordinal scale. The rank order of proximities determines the dimensionality of the space and the metric configuration of the points representing the stimuli (Shinkareva et al., 2013). As reported in previous studies in this research field, we assumed this space to be bidimensional, with valence and arousal as orthogonal dimensions (Russell & Bullock, 1985). The goodness-of-fit the MDS representation is estimated with the Stress measure.

Analysis of emotional (valence and arousal) and visual complexity ratings. Valence and arousal ratings were entered as dependent variables in two repeated-measures ANOVAs, with picture type (emotional vs neutral) as a within-group factor in experiment 1, and category as a within-group factor in experiment 2-3. Moreover, we analysed the visual complexity ratings from the fMRI task by transforming them into a continuous variable. Specifically, for each subject we calculated the proportion of 'high complexity' responses, by dividing the number of 'high complexity' responses within each category by 18 (the number of pictures within each category), and then averaged them across sessions. These were entered as dependent variables in a repeated-measure ANOVA, with category

(i.e., E1, E2, N1, and N2) as grouping factor. The results from the valence and arousal ratings of experiment 1 are reported in Table 4.1, those from experiment 2-3 in Table 4.2. The results of the visual complexity rating task are shown in Table 4.4. Data analyses were conducted in Matlab R2018 (MATLAB 2018a, The MathWorks, Inc., Natick, Massachusetts, United States), and SPSS (IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp).

Categories				**Statistics**		
E1	**E2**	**N1**	**N2**	**F**	**p**	**η_p^2**
.308 ± .163	.728 ± .211	.297 ± .187	.267 ± .170	5.34	<.001**	0.63
Post hoc						
E1 vs E2	**E1 vs N1**	**E1 vs N2**	**E2 vs N1**	**E2 vs N2**	**N1vs N2**	
-.42, .000**	.01,.1.00	.04,.1.00	.43,.000**	.46,.000**	.03,1.00	

Table 4.4. Differences in visual complexity ratings among categories. The proportion of high complexity ratings within each category (total number of 'high complexity' responses divided by 18) was averaged across sessions. Mean and standard deviation of each category, and the statistics of the difference among them are reported at the top of the table. Bonferroni post hoc corrections for multiple comparisons ($p<0.05$) are summarized at the bottom. *, pFWE< 0.05; **, pFWE< 0.001.

Neuroimaging data analysis

Preprocessing.

Neuroimaging data were pre-processed and analysed using Statistical Parametric Mapping (SPM12) (http://store.elsevier.com/product.jsp?isbn=9780123725608) and MATLAB R2018a (MATLAB 2018a, The MathWorks, Inc., Natick, Massachusetts, United States). Functional images were slice-time corrected to reduce the mismatching between acquisition timing of different slices, and realigned to a reference (mean) image to minimize the variance due to head movements. These were then coregistered to the high-resolution T1-weighted structural image, which was coregistered and normalized to MNI space. Finally, functional images were normalized to a standard template volume based on the Montreal Neurological Institute (MNI) reference brain to achieve a more precise comparison across individuals. Spatial smoothing was performed only on functional data analysed with a conventional univariate approach using a 6-mm full-width at half-maximum isotropic Gaussian kernel. No spatial smoothing was carried over on the multivariate functional data, according to the standard practices for MVPA studies (Haxby et al., 2001; Kriegeskorte, Mur, et al., 2008b). The preprocessing for the univariate tests was identical to the one for the RSA with the

exception of using a 6-mm FWHM Gaussian smoothing kernel (Kriegeskorte et al., 2006).

Individual-level model for RSA analysis

After preprocessing, functional data from each voxel were analysed using the general linear model (GLM). Each stimulus was modelled as a separate event beginning with picture presentation onset, using the canonical function in SPM12, and included in the model as regressor of interest (72 regressors per session). Six motion correction parameters were also modelled within each session, and included in the model as regressor of no interest. From this GLM analysis, we obtained a single beta image for each stimulus. Contrast images for each stimulus against the implicit baseline were generated based on the fitted responses, and averaged across sessions. The resulting 72 T-contrast images were used as inputs for RSA.

Individual-level models for univariate analyses

Although our hypotheses were specific to the multivariate representations, we also performed three conventional univariate analyses, referred to as GLM1, 2 and 3. GLM1 was performed as a manipulation check, to evaluate the probability that any differences in the RSA analysis were due to differences in the average univariate activations among conditions. For this reason, GLM1 used individual-level models that were almost identical to those used for the RSA, the only difference being that instead of modelling 72 stimuli, here each category (i.e., E1, E2, N1, N2) was modelled as separate condition (4 regressors per session) beginning with each picture presentation onset, using the canonical function in SPM12.

GLMs 2-3 were performed as a second manipulation check, to test whether our study replicated previous findings showing higher recruitment of emotional regions during the processing of emotional than neutral stimuli, across the 4 sessions (GLM 2) and within session 1 only (GLM 3). For this reason, individual-level models were altered to be maximally sensitive to the difference between emotional and neutral stimuli. Specifically, in GLM 2-3 we included the temporal derivative, to take into account temporal differences in the BOLD signal between emotional and neutral conditions (Calhoun, Stevens, Pearlson, & Kiehl, 2004; Friston et al.,

1998; Heinzel et al., 2005). GLM3 was performed on session1 data only to check whether any ROI was significantly more activated in session 1 only and then decreased its activation after further repetition.

ROIs definition.

We defined the ROIs by using the Automated Anatomical Labelling (AAL) template in WFU Pickatlas toolbox (https://www.nitrc.org/projects/wfu_pickatlas) and Anatomy toolbox (https://www.fil.ion.ucl.ac.uk/spm/ext/#AAL), and constructed with MarsBaR 0.43 (http://marsbar.sourceforge.net). We used WFU_Pickatlas toolbox to define the bilateral early visual cortex (EVC) as Broadmann (Ba) 17, the Dorsomedial Prefrontal cortex (DMPFC) corresponded to the Ba 8 and 9, the Ventromedial Prefrontal cortex (VMPFC) to the Ba 10, and the (dorsal and ventral) Anterior Cingulate Cortex (ACC) to the Ba 32 and 24.The Retrosplenial cortex (RSC), the occipital place area (OPA) and the Parahippocampal place area (PPA) were respectively defined: the bilateral RSC as Ba 29 and Ba 30; the OPA as an 8 mm sphere around the coordinates reported by Julian et al. (2016) (Julian, Ryan, Hamilton, & Epstein, 2016) (left OPA: -34, -77, 21; right OPA: 34, -77, 21) (Julian et al., 2016); the PPA as an 8 mm sphere around the coordinates reported by (Henson & Mouchlianitis, 2007) (left PPA: -27, -45, -12; right PPA 30, -42, -9). The Face Fusiform Area (FFA) was defined as an 8 mm sphere around the coordinates reported by (Henson & Mouchlianitis, 2007) (left FFA: -42, -51, -18; right FFA: 42, -45, -21). The medial Temporal lobe (MeTL) comprised the Entorhinal cortex defined with Anatomy toolbox, and the bilateral Hippocampus, the Perirhinal cortex, the Parahippocampal cortex defined with AAL. The same toolbox was used for the bilateral Inferior Temporal cortex (ITC), the anterior temporal lobe (ATL), the Amygdala, the Thalamus, the Insula, the Precuneus and the bilateral Orbitofrontal cortex (superior, middle, inferior and medial OFC). We combined these ROIs into one 'ROIs mask', which was used in the searchlight RSA.

Univariate group analyses

From each individual-level GLM, we obtained a single beta image for each condition. We then compared emotional and neutral conditions (emotional > neutral), thereby producing one contrasted image for each subject. The contrasted image from each subject was then entered as dependent variable in a one sample

t-test. Both the univariate and the multivariate results were inclusively masked to only include our Regions of Interest (ROIs) involved in the visual, semantic and emotional processing of complex pictures, as defined in the paragraph about ROIs definition.

RSA group analyses: quantifying neural similarity

Brain- behaviour correlations. In order to test our main hypothesis (i.e., higher neural similarity between the two emotional than the two neutral categories), we first conducted a very precise localisation technique, the searchlight RSA, to investigate which brain regions (within the ROIs mask) represented the participants' similarity space. This was carried out by computing the Spearman's correlation between brain activation-patterns RDMs and behavioural RDMs (second order isomorphism). The behavioural RDM represented the participants' similarity space resulted from the MA task, created as explained in the paragraph about the behavioural data analysis. Three separate analyses were conducted. The first used the entire RDM (with all the 72 stimuli, 'all RDM'); the second focused exclusively on the emotional stimuli (36 stimuli, 'emotional RDM'), and the third on the neutral stimuli (36 stimuli, 'neutral RDM'), depicted as violet and green squares at the bottom of Figure 4.1, respectively. We conducted these latter two analyses to explore whether any brain region was involved in the representation of either the emotional or the neutral categories. For the purpose of these three analyses, three brain activation-pattern RDMs were constructed for each participant in the same way. The participant's brain activation-pattern RDMs were computed by entering the T-contrast images into a matrix with all the voxels in the rows, and the experimental stimuli in the columns. Then, for each subject and each of the three analyses, 3 × 3 × 3 voxels spherical cluster was moved throughout the brain and at each location in the ROIs mask a correlational distance (among T values) was assigned to the centre voxel of the sphere, resulting in a (x, y, z, number of pairs) brain activation patterns RDM for each subject. This measure quantified the dissimilarity across voxels in a given searchlight sphere for each specific pair. The number of pairs represented all the possible combinations between experimental stimuli (2556 pairs with 72 stimuli, 630 with 36 stimuli). Next, for each stimulus, the similarity between brain and behavioural RDMs was estimated using a pairwise Spearman's correlation. This provides a correlational

map between the behavioural and the brain RDMs for each subject, which reveals where the similarity space is best represented in the brain (highest correlation), and an ‘n map’, wherein the number of voxels that contributes to each correlational value is reported in each entry. The correlational coefficients were Fisher's z transformed, and inference was performed at each voxel by performing a one side signed rank test across subjects, testing the null hypothesis of no correlation between brain and behaviour RDMs. The resulting p values (uncorrected) were thresholded to control the false-discovery rate (FDR). We performed two different FDR correction procedures, to yield a more conservative as well as a more lenient set of results. In both procedure, the p-values (uncorrected) were first ordered such that $p1 \leq p2 \leq pn$. Then, in the most conservative procedure we divided the rank number (of each p-values) by the total number of voxels in the ROIs mask and multiplied it for alpha. In the more lenient procedure, we divided the rank number (of each p-values) by the number of voxels that contribute to each correlational value (between brain and model RDM), and multiplied it for alpha. The number of voxels was extracted from the n map associated with the correlational map.

Differences in neural dissimilarity between emotional and neutral categories. We conducted a second set of analyses to test our main hypothesis, that is, higher neural similarity (lower dissimilarity) between the two emotional compared to the two neutrals categories (similarity E12>N12). We tested this effect in the brain clusters that we observed to be involved in representing both the entire (72 stimuli) and partial (36 stimuli) participants’ similarity space. With this aim, we created different masks, one for each significant cluster. In case of ROIs that were significantly correlated with both the emotional and the neutral similarity space, we selected the clusters correlated with the neutral similarity space. Then, for each subject and each mask, we computed a brain activation-pattern RDM, where each entry represented the correlational distance (1- Spearman’s correlation) between brain activations across voxels within that mask, and the rows and the columns represented the experimental stimuli. We will refer to this as ROI RDM. It is symmetric about a diagonal of zeros, and resulted in 2556 cells in the lower triangular part that reflected the pairwise dissimilarity of the response patterns associated with the stimuli for each ROI. Then, within each participant ROI RDM,

we calculated the mean of the conditions of interest (E12 and N12), and entered them as dependent variables in paired t- tests, one for each cluster ($p<0.05$). The RSA was performed using the MRC-CBU RSA toolbox for MATLAB (http://www.mrc-cbu.cam.ac.uk/methods-and-resources/toolboxes).

As in the behavioural experiments, we tested any differences in the variance across participants in the neural dissimilarity between E12 and N12. We explored this effect in brain clusters wherein we observed significant differences in neural dissimilarity between E12 and N12. We conducted different two-samples F-tests for variance, one for each cluster.

4.3 Results

In a series of experiments with different datasets of real-world pictures, we explored whether emotions are associated with increased perceived similarity, both subjective (ratings) and objective (neural) similarity. We hypothesised that for both dependent measures, perceived similarity will be higher (dissimilarity lower) (I) within category, compared to between categories, and (II) between the emotional categories compared to the neutral categories.

Behavioural evidence for increased similarity between emotional stimuli

Experiment 1 confirmed our hypotheses. We observed a significant main effect of our conditions ($F(2, 18) = 91.00$, $p<0.001$, $\eta_p^2 = 0.83$), with lower dissimilarity within (i.e., EE and NN) than between (i.e., EN) categories ($p<0.001$), and in EE than in NN ($p<0.001$). When represented in the similarity space, emotional pictures were displaced closer to each other than neutral pictures. This resulted in a Stress value of 0.05, indicating a good fit of this model. These findings are shown in Figure 4.2.

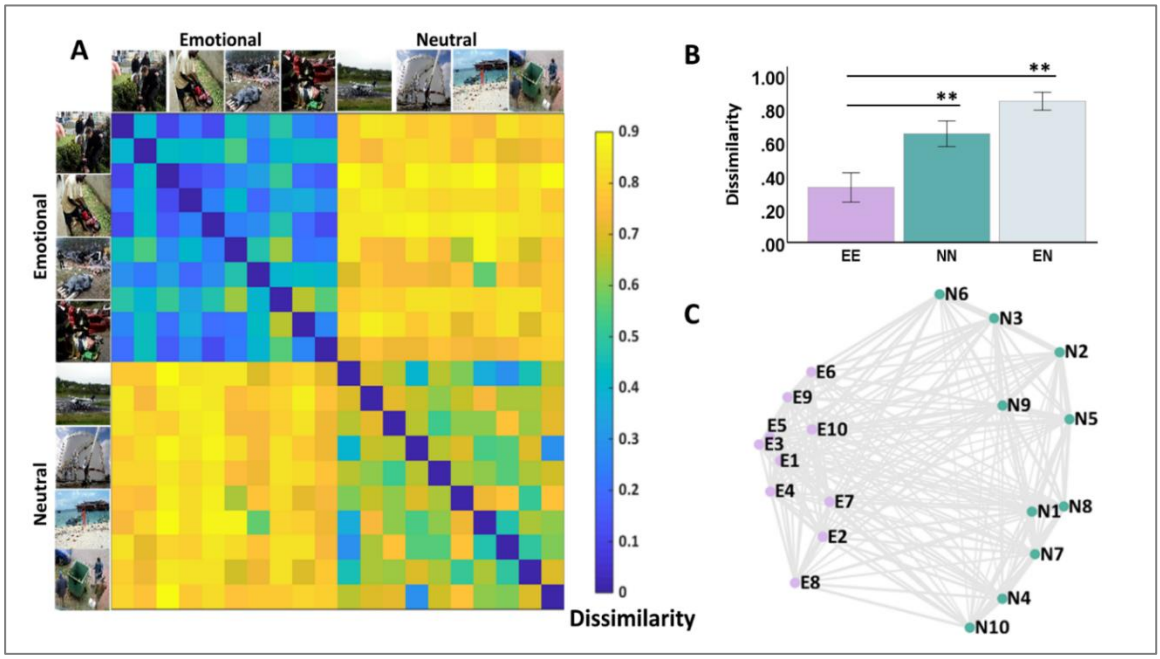


Figure 4.2. A) Representational Dissimilarity Matrix (RDM) of 20 complex pictures (10 emotional, 10 neutral), averaged across participants. It is symmetric about a diagonal of zeros, the rows and the columns represent the stimuli, and each cell the dissimilarity, measured as 1- standardized similarity ratings between stimuli within each specific pair. Yellow colours denote high dissimilarity, blue colours low dissimilarity. B) The average dissimilarity within emotional pictures (EE), within neutral pictures (NN), and between emotional and neutral pictures (EN, grey). Error bars represent ±2 SEM; **p<0.001. C) The Multidimensional Scaling (MDS) plot of the 20 pictures in a bidimensional space. Additional information supporting Figure 4.2 can be found in Figure 2-1.

In experiment 2, with the second dataset, which controlled for the higher thematic similarity between emotional pictures, we observed different results. Specifically, we found lower dissimilarity in E12 and N12 compared to EN ($F(2, 38) = 27.40$, $p<0.001$, $\eta_p^2= 0.41$), but no differences in similarity ratings between the two emotional and the two neutral categories. The same results were replicated using the MA task in experiment 3. Our manipulation check revealed lower dissimilarity within category (i.e., EE and NN) than between categories (i.e., E12, N12, EN) ($F(4, 26) = 214.76$, $p<0.001$, $\eta_p^2= 0.88$), but no difference due to emotion in the critical comparison between E12 and N12. In the bidimensional space, the proximities between the two emotional, and between the two neutral categories do not differ. The Stress value was 0.10, indicating a fair fit of this model. This reduction in the goodness of fit compared to experiment 1 might suggest that the weight of the semantic dimension in subjective similarity may have been higher in experiments 2-3, where four categories were included, compared to experiment 1, where stimuli were not grouped by semantic category. These findings are shown in Figure 4.3.

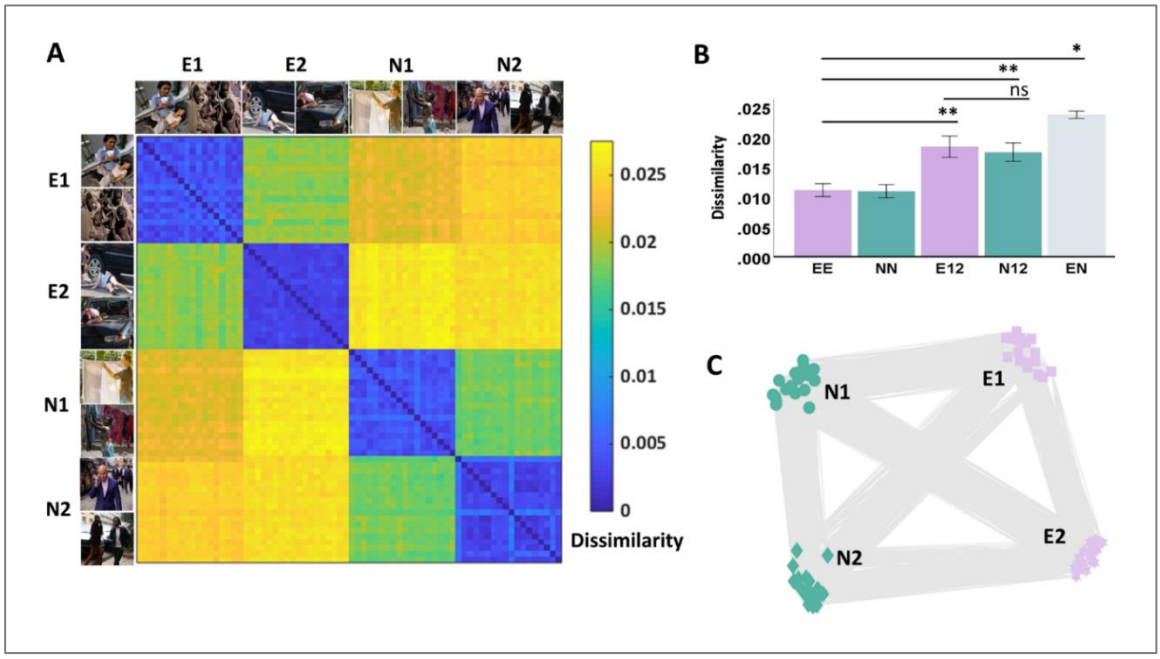


Figure 4.3. A) Representational Dissimilarity Matrix (RDM) of 72 complex pictures (Emotional categories: E1, poverty (1 to 18); E2, car accidents (19 to 36); Neutral categories: N1, laundry (37 to 54); N2, phone call (55 to 72), averaged across participants. It is symmetric about a diagonal of zeros, the rows and the columns represent the stimuli, and each cell the dissimilarity (measured as Euclidean distance) between stimuli within each specific pair. Yellow colours denote high dissimilarity, blue colours low dissimilarity. B) The average dissimilarity within emotional pictures (averaged across E1 and E2) (EE), within neutral pictures (averaged across N1 and N2) (NN), between emotional pictures (E12), between neutral pictures (N12), and between emotional and neutral pictures (EN). Error bars represent ±2 SEM; *, pFWE<0.05; **, pFWE<0.001. C) The Multidimensional Scaling (MDS) plot of the 72 pictures in a bidimensional space. Additional information supporting Figure 4.3 can be found in Figure 3-1.

Finally, in all the experiments, we did not observe any significant differences in the variance across participants between emotional and neutral conditions. This allows us to exclude an alternative explanation of the behavioural results, that is, that the similarity between emotional pictures can be affected by individuals' emotional granularity (Barrett, Gross, Christensen, & Benvenuto, 2001). High-granular individuals would be more aware of the differences of their emotional experiences when viewing pictures from the two emotional categories and may rate them as less similar, while low-granular individuals may rate them as more similar, ultimately masking the difference between emotional and neutral categories. However, if this explanation was correct, we would expect increased variance in ratings of emotional pictures. Instead, there were no significant differences in rating variance between emotional and neutral categories. These results are reported in Table 4.5.

Behavioral experiments	Conditions		F value	F critical	P value
Experiment 1	EE 0.04	NN 0.03	1.28	2.53	0.36
Experiment 2	E12	N12	0.72	1.89	0.16

	0.05	0.07			
Experiment 3	EE 0.000	NN 0.000	0.79	2.13	1.00
	E12 0.000	N12 0.000	1.34	2.13	0.44

Table 4.5. Differences in the variance in similarity judgements between emotional and neutral stimuli. The variance averaged across participants for each conditions, and the statistics of each difference between conditions are reported. In experiment 1, EE and NN represent the variance within emotional, and within neutral stimuli, respectively, averaged across participants. In experiment 2-3, E12 and N12 signify the variance between the two emotional, and the two neutral categories, respectively, averaged across participants. Finally, in experiment 3, EE and NN represent the variance within E1 and E2, and within N1 and N2, averaged across participants.

Manipulation check: Univariate differences between the emotional and neutral conditions

In GLM 1, no clusters (number of voxels > 10) survived the correction for multiple comparisons, suggesting that RSA results are unlikely to be contaminated by mean signal differences. Conversely, we replicated previous findings in GLM 2-3; these results are reported in Table 4.6 and Figure 4.4.

Analysis	p FWE	K	p uncorr	T	x	y	z	Label
GLM2 Emo>Neu, all sessions	0.001	197	<0.001	8.33	36	-46	-19	FFA R
	0.023	114	0.002	8.16	-39	-46	-19	FFA L
	0.039	100	0.003	6.72	-39	-82	11	OPA L
	0.003	178	<0.001	5.49	33	26	-4	Insula R
	<0.001	107	<0.001	4.89	-27	23	-4	Insula L
GLM3 Emo>Neu Session1	0.014	104	0.001	6.92	6	-13	-1	Thalamus R
	0.029	88	0.002	6.24	-39	-46	-19	FFA L
	0.003	138	<0.001	5.65	39	-64	-10	Temporal inferior R/ FFA R
	0.002	146	<0.001	5.58	-36	20	2	Insula L
	0.038	82	0.002	5.26	-39	-82	11	OPA L
	0.026	90	0.002	4.59	48	35	-10	OFC R

Table 4.6. Differences in BOLD signal change between emotional and neutral categories. Only regions that survive correction for multiple comparisons using pFWE < 0.05 are reported. Small volume correction using the ROI mask was applied in both analyses.

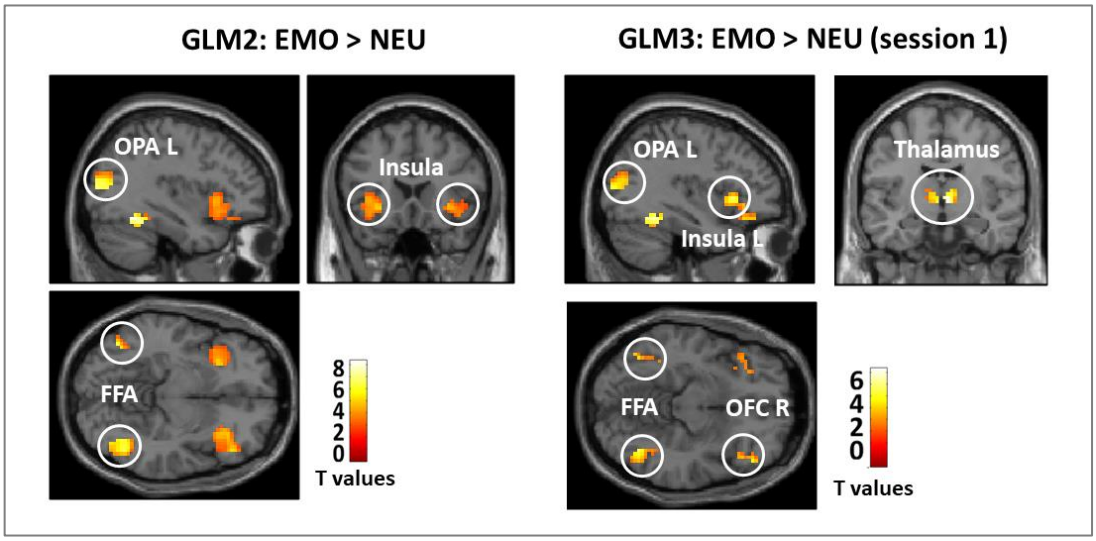


Figure 4.4. Differences in BOLD signal change between emotional and neutral categories, across 4 sessions (GLM 2, left) and in session 1 only (GLM 3, right). Only regions that survive correction for multiple comparisons using pFWE < 0.05 are reported. Small volume correction using the ROI mask was applied in both analyses.

Brain- behaviour correlations

We carried out a searchlight RSA to investigate the brain regions within the ROIs mask that represented the participants' self-reported similarity space. First, we tested whether the neural-pattern similarity within the ROIs mask was significantly correlated with the entire (72 x 72) similarity space, comprised of neutral and emotional categories. These data only survived our more lenient correction for multiple comparisons (p_{FDR}<0.05) (see Materials and Methods). We observed that clusters in the bilateral ITC, the right FFA, and the right Precuneus represented the participants' similarity space. These findings are reported in in Table 4.7 and Figure 4.5A.

Analysis	Regions	x	y	z	n voxels	t	p_{FDR}	d
ALL RDM (72 x 72)	**ITC L**	11	-61	-9	154	38.81	<0.001**	3
	ITC R	46	-63	-9	21	40.5	<0.001**	10
	FFA R	42	-47	-19	69	29.64	<0.001**	4
	Prec R	3	-57	20	32	13.11	<0.001**	3
Emotional RDM (36 x 36)	**EVC L**	-13	-86	4	72	15.1	<0.001**	1.82
	EVC R	6	-93	-1	21	16.17	<0.001**	4.28
	OPA L	-34	-80	18	103	117.25	<0.001**	11.43
	OPA R	34	-80	18	103	74.97	<0.001**	7
	PPA L	-21	-43	-10	32	58.57	<0.001**	11.25
	PPA R	24	-43	-16	29	89.19	<0.001**	16
	FFA L	-36	-49	-22	41	25.38	<0.001**	5
	FFA R	39	-49	-22	34	26.18	<0.001**	4
	Prec	-2	-60	40	1249	109.58	<0.001**	4
	dACC L	-3	27	1	18	27.79	<0.001**	2.5

	dACC R	3	15	42	20	29.18	<0.001**	7.5
	aINS L	-35	27	1	18	27.79	<0.001**	7.5
Neutral RDM (36 x 36)	**OPA L**	-29	-78	20	11	20.44	<0.001**	6.67
	OPA R	36	-79	16	43	57.2	<0.001**	10
	PPA L	-30	-49	-10	17	18.7	<0.001**	4
	PPA R	25	-46	-13	10	7.19	<0.001**	4.28
	FFA L	-42	-52	-13	10	7.19	<0.001**	2.25

Table 4.7. Brain-behaviour correlations. Top: correlations between the entire (72 x 72) stimulus space (named as 'all RDM'), the and the brain. Significant correlations were observed in the bilateral ITC, right FFA, and the right Prec. Middle: correlations between the emotional (36 x 36) similarity space (named as 'emotional RDM') and the brain. Significant correlations were observed in the bilateral EVC, OPA, PPA, FFA, Prec, dACC and left aINS. Bottom: correlations between the neutral (36 x 36) similarity space (named as 'neutral RDM') and the brain. Significant correlations were observed in the bilateral OPA, PPA, and left FFA. In all these analyses, correlational coefficients were Fisher's z transformed, and entered as dependent variables in a one side t test (separately for each brain region), testing the null hypothesis of no correlation between the participants' similarity space and the neural activation patterns. The resulting p values were thresholded to control for the false-discovery rate (FDR). **, pFDR< 0.001. Abbreviations. ITC, Inferior Temporal Cortex; FFA, Face Fusiform Area; Prec, Precuneus; EVC, Early visual cortex; OPA, Occipital place area; PPA, Parahippocampal place area; dACC, dorsal anterior cingulate cortex; aINS, anterior insula. L, Left; R, Right.

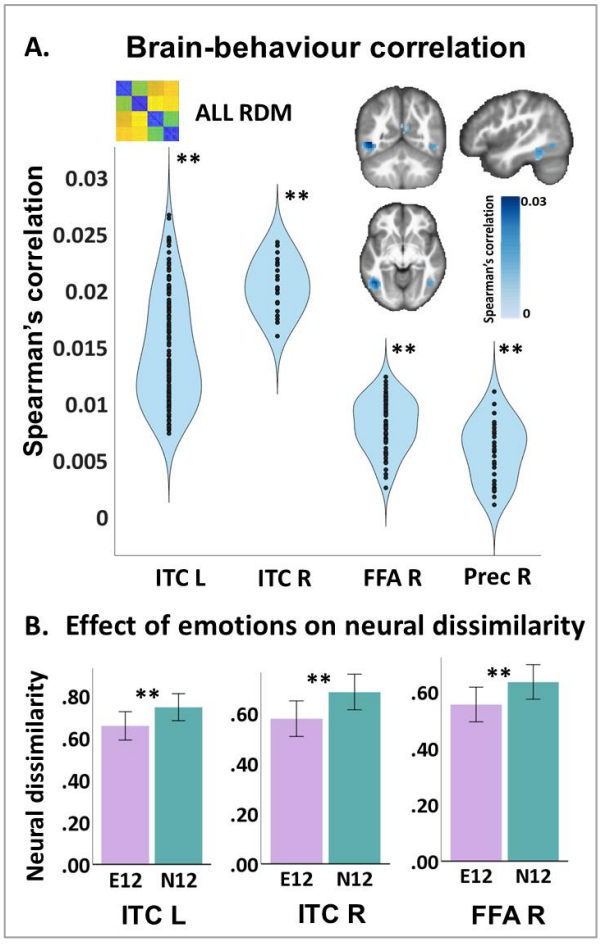


Figure 4.5. A) Correlation between the entire (72 x 72) stimulus space (named as 'all RDM') and the brain. Significant correlations were observed between the behavioural 'all RDM' and clusters in the bilateral ITC, right FFA, and the right Prec. Correlational coefficients were Fisher's z transformed, and entered as dependent variables in a one side t test (separately for each brain region), testing the null hypothesis of no correlation between the participants' similarity space and the neural activation patterns. The resulting p values were thresholded to control for the false-discovery rate (FDR). **, pFDR< 0.001. B) Differences in neural dissimilarity (measured as correlational distance) between emotional and neutral stimuli in different brain clusters, including the bilateral ITL, and the right FFA. The dissimilarity between emotional categories (E12) was calculated by averaging the dissimilarity between E1 and E2, and the dissimilarity between neutral

categories (N12) by averaging the dissimilarity between N1 and N2, for each participant. These were entered as dependent variables in paired t tests, one for each brain cluster ($p<0.05$). **, $p< 0.001$. Abbreviations. ITC, Inferior Temporal Cortex; FFA, Face Fusiform Area; L, Left; R, Right.

Second, we performed the same analysis separately for the emotional and neutral pictures to explore whether any brain region was involved in the representation of either the emotional or the neutral categories (see the violet and green squares in Figure 4.1). The results from these analyses survived the more conservative correction for multiple comparisons ($p_{FDR}<0.05$) (see Materials and Methods). We found that participants' emotional similarity space was significantly correlated with clusters in lower and higher-level visual processing regions, as well as regions involved in emotional processing. These included the bilateral EVC, bilateral OPA, bilateral PPA, bilateral FFA, bilateral Precuneus, bilateral dorsal ACC and in the left anterior Insula. By contrast, participants' neutral similarity space was significantly correlated with clusters in higher-level visual regions only, including the bilateral OPA, the bilateral PPA, and left FFA. These findings are reported in Table 4.7 and Figure 4.6 A and 4.7A.

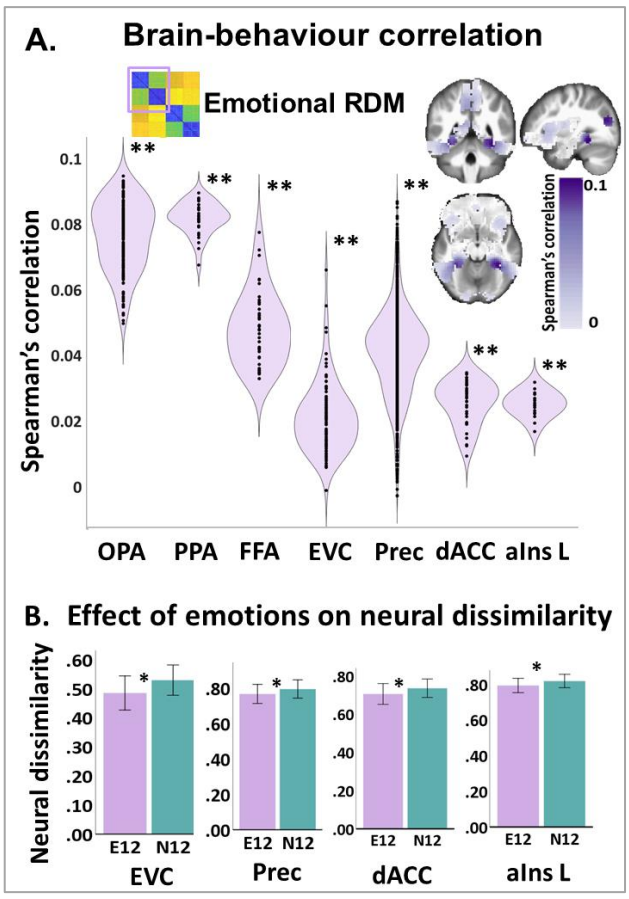


Figure 4.6. A) Correlation between the emotional (36 x 36) similarity space (named as 'emotional RDM') and the brain. Significant correlations were observed between the behavioural 'emotional RDM' and clusters in the bilateral OPA, PPA, FFA, EVC, Prec, dACC, and left aIns. Correlational coefficients were Fisher's z

transformed, and entered as dependent variables in a one side t test (separately for each brain region). For simplicity, we averaged the left and the right sides of the clusters wherein both sides were significant. The resulting p values were thresholded to control the false-discovery rate (FDR). **, pFDR< 0.001. B) Differences in neural dissimilarity (measured as correlational distance) between emotional and neutral stimuli in different brain clusters, including the bilateral EVC, Prec, dACC and left aIns. The dissimilarity between emotional categories (E12) was calculated by averaging the dissimilarity between E1 and E2, and the dissimilarity between neutral categories (N12) by averaging the dissimilarity between N1 and N2, for each participant. These were entered as dependent variables in paired t tests, one for each brain cluster (p<0.05). *, p< 0.05. Abbreviations. OPA, Occipital place area; PPA, Parahippocampal place area; FFA, Face fusiform area; EVC, Early visual cortex; Prec, Precuneus; dACC, Dorsal anterior cingulate cortex; aIns, Anterior insula; L, left; E12, dissimilarity between emotional categories; N12, dissimilarity between neutral categories.

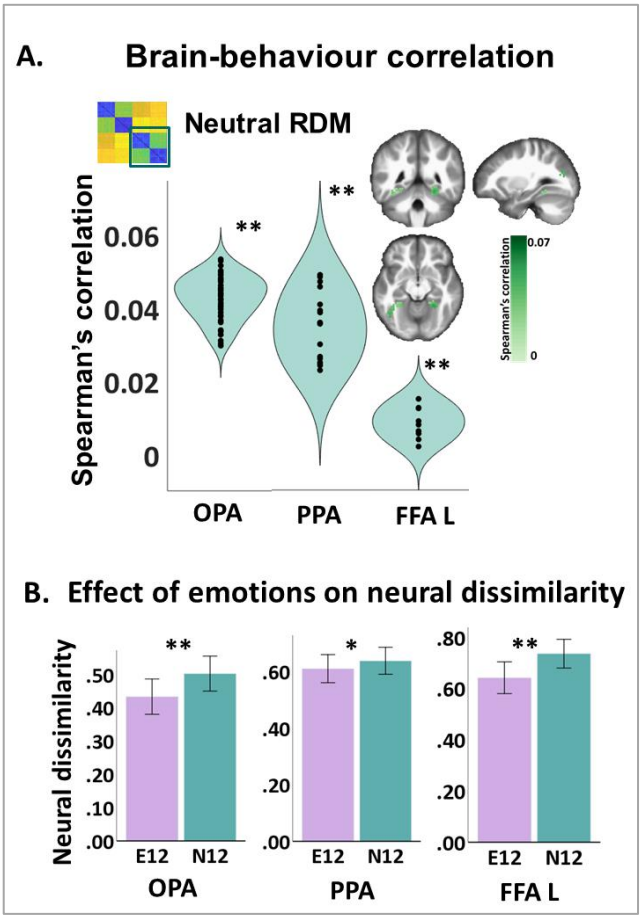


Figure 4.7. A) Correlation between the neutral (36 x 36) similarity space (named as 'neutral RDM') and the brain. Significant correlations were observed between the behavioural 'neutral RDM' and clusters in in the bilateral OPA, PPA and left FFA. Correlational coefficients were Fisher's z transformed, and entered as dependent variables in a one side t test (separately for each brain region). For simplicity, we averaged the left and the right sides of the clusters when both sides were significant. The resulting p values were thresholded to control the false-discovery rate (FDR). *, pFDR< 0.05; **, pFDR< 0.001. B) Differences in neural dissimilarity (measured as correlational distance) between emotional and neutral stimuli in different brain clusters, including the bilateral OPA, PPA and left FFA. The dissimilarity between emotional categories (E12) was calculated by averaging the dissimilarity between E1 and E2, and the dissimilarity between neutral categories (N12) by averaging the dissimilarity between N1 and N2, for each participant. These were entered as dependent variables in paired t tests, one for each brain cluster (p<0.05). *, p< 0.05; **, p< 0.001. Abbreviations. OPA, Occipital place area; PPA, Parahippocampal place area; FFA, Face fusiform area; L, left; E12, dissimilarity between emotional categories; N12, dissimilarity between neutral categories.

Neural evidence for increased similarity between emotional stimuli

We performed the ROIs RSA to explore whether the neural representations of emotional categories are more similar than those associated with neutral

categories. This analysis was carried out in brain clusters from the above analysis, namely, those that significantly correlated with the whole participants' similarity space (Figure 4.5B), as well as with its emotional (Figure 4.6B), and neutral (Figure 4.7B) similarity spaces. As predicted, the neural pattern dissimilarity of emotional categories was lower than the one of neutral stimuli in all the previously reported clusters ($p<0.05$), apart from the right PPA. In addition, we observed trends towards significance in support to our hypothesis in the right EVC ($p=0.11$) and in one cluster in the left PPA ($p=0.06$). These findings are reported in Table 4.8 and in Figures 4.5B, 4.6B and 4.7B.

Analysis	ROIs	E12	N12	t	p	d
ALL RDM (72 x 72)	**ITC L**	0.66 ± 0.18	0.75 ± 0.17	-7.90	<0.001**	-0.51
	ITC R	0.58 ± 0.19	0.69 ± 0.19	-10.88	<0.001**	-0.58
	FFA R	0.56 ± 0.17	0.64 ± 0.17	-8.89	<0.001**	-0.47
	Prec R	0.80 ± 0.10	0.83 ± 0.09	-3.60	0.001*	-0.32
Emotional RDM (36 x 36)	**EVC L**	0.57 ± 0.19	0.64 ± 0.17	-6.20	<0.001**	-0.39
	EVC R	0.49± 0.16	0.53 ± 0.14	-1.65	0.110	-0.27
	OPA L	0.49 ± 0.18	0.55 ± 0.17	-4.69	<0.001**	-0.34
	OPA R	0.42 ± 0.17	0.48 ± 0.17	-5.36	<0.001**	-0.35
	PPA L	0.83 ± 0.09	0.85 ± 0.04	-1.93	0.064	-0.31
	PPA R	0.64 ± 0.14	0.64 ± 0.14	-0.64	0.523	0.00
	FFA L	0.64 ± 0.17	0.74 ± 0.15	-11.34	<0.001**	-0.63
	FFA R	0.55 ± 0.17	0.63 ± 0.16	-8.88	<0.001**	-0.48
	Prec	0.77 ± 0.15	0.80 ± 0.14	-3.56	0.001*	-0.21
	dACC L	0.68 ± 0.15	0.71 ± 0.14	-3.19	0.003*	-0.21
	dACC R	0.73 ± 0.16	0.76 ± 0.14	-2.40	0.023*	-0.20
	aINS L	0.79 ± 0.11	0.82 ± 0.10	-2.85	0.008*	-0.29
Neutral RDM (36 x 36)	**OPA L**	0.47 ± 0.17	0.54 ± 0.16	-7.25	<0.001**	-0.42
	OPA R	0.40 ± 0.15	0.46 ± 0.16	-5.86	<0.001**	-0.39
	PPA L	0.55 ± 0.13	0.58 ± 0.14	-2.79	0.009*	-0.22
	PPA R	0.47 ± 0.13	0.47 ± 0.14	-0.73	0.473	0.00
	FFA L	0.64 ± 0.17	0.77 ± 0.15	-11.34	<0.001**	-0.81

Table 4.8. Effect of emotions on neural dissimilarity. Difference in neural dissimilarity (measured as correlational distance) among conditions. The dissimilarity between emotional categories (E12) was calculated by averaging the dissimilarity between E1 and E2, and the dissimilarity between neutral categories (N12) by averaging the dissimilarity between N1 and N2, for each participant. These measures were first computed in brain clusters significantly involved in the representation of the whole (72 stimuli) participants' similarity space (top of the table). Then, we computed E12 and N12 in brain clusters significantly involved in the representation of the emotional (middle of the table) and neutral (bottom of the table) participants' similarity space. We entered them as dependent variables in paired t tests, one for each brain cluster. Bonferroni post hoc corrections for multiple comparisons ($p<0.05$) are summarized at the bottom. *, pFWE< 0.05; **, pFWE< 0.001. Abbreviations. E12, neural dissimilarity between emotional categories; N12, neural dissimilarity between neutral categories. ITC, Inferior temporal cortex; FFA, Face fusiform area; Prec, Precuneus; EVC, Early visual cortex; OPA, Occipital place area; PPA, Parahippocampal place area; dACC, Dorsal anterior cingulate cortex; AI, Anterior insula; L, left; R, Right.

Finally, we did not observe any significant differences in the variance across participants between E12 and N12 in any brain clusters. These results are reported in Table 4.9.

ROIs	E12	N12	F value	F critical	P value
ITC	0.03`	0.03	1.04	2.13	0.45
Precuneus	0.02	0.02	1.09	2.13	0.82
EVC	0.02	0.02	1.27	2.13	0.52
OPA	0.02	0.02	1.02	2.13	0.97
PPA	0.02	0.02	0.99	2.13	1.00
FFA	0.02	0.02	1.07	2.13	0.85
dACC	0.02	0.02	1.28	2.13	0.52
aIns L	0.01	0.01	1.15	2.13	0.71

Table 4.9. Differences across participants in the variance in neural dissimilarity between emotional and neutral stimuli. The variance averaged across participants for each conditions within each cluster, and the statistics of each difference between conditions are shown. E12 and N12 represent the variance between the two emotional, and the two neutral categories, respectively, averaged across participants. For simplicity, we averaged the left and the right sides of the clusters. Abbreviations. ITC, Inferior Temporal Cortex; EVC, Early visual cortex; OPA, Occipital place area; PPA, Parahippocampal place area; FFA, Face fusiform area; Prec, Precuneus; dACC, Dorsal anterior cingulate cortex; aIns, Anterior insula; L, left; E12, dissimilarity between emotional categories; N12, dissimilarity between neutral categories.

4.4 Discussion

We investigated behavioural and neural similarity measures between complex emotional and neutral stimuli using two similarity judgements tasks and two stimulus databases, the second of which was very tightly controlled. We report two novel findings. First, the similarity between neural representations of stimuli from two negatively-valenced, emotionally-arousing categories was greater than the neural similarity between stimuli from two neutral categories. This increase was observed while participants were processing individual stimuli rather than inter-stimulus relationships. Some, but not all, of the clusters expressing similarity among emotional stimuli preferentially also expressed similarity among neutral stimuli. Second, once semantic similarity was controlled, participants rated the similarity of stimuli from two emotional categories to be equivalent to that of stimuli from two neutral categories. Thus, the greater neural similarity between emotional pictures did not influence perceived similarity in the same participants. We discuss the implications of these results below.

Increased neural similarity between emotional than neutral realistic events.

In experiment 3, we observed increased neural similarity between emotional than neutral categories, such that in brain clusters involved in encoding participants' entire similarity space, the neural similarity between emotional categories was stronger than between neutral categories. These clusters were located in the ventral visual stream, which underpins semantic categorisation (Clarke & Tyler, 2014), and in regions involved in affect representations (e.g., precuneus) (Kim,

Shinkareva, & Wedell, 2017) and modulation (e.g., dACC)(Saarimäki et al., 2018). To our knowledge, this is the first report of the neural underpinnings of perceived similarity between complex emotional stimuli, while using a pictures set controlled for visual and semantic attributes.

This finding has implications for research about the neurobiological correlates of categorisation and generalisation. Previous studies (Dunsmoor et al., 2013; R. M. Visser et al., 2011) observed increased neural similarity among exemplars that predicted threat. They proposed that this mechanism was adaptive, enabling individuals to differentiate emotionally-salient stimuli from others, and supporting broad generalisation between items which predict fitness-relevant outcomes. Although our work differed from these studies, where the emotional response was induced though Pavlovian conditioning, we found the same effect here. This converging evidence suggests that it is evolutionarily more important to integrate emotional information in neural representations to increase the relevance and generalisability of stimuli that predict a negative outcome. These findings concur with the conclusions that emotion serves as a fundamental feature of cognition, such that any representation of the world is an integrated product between emotion, perception and thought (e.g., “That is a good thing”) rather than discrete and isolated psychological events (e.g., “That is a thing. I feel good”) (Todd et al., 2020).

We extended previous findings about brain regions involved in representing emotional categories and dimensions by exploring, for the first time, differences in the neural representations of the relationships between emotional and neutral stimuli. The bilateral ITC, right FFA and right precuneus represented the entire similarity space and exhibited greater neural similarity between the two emotional than the two neutral categories. As part of the hierarchical network in the ventral visual stream, the ITC integrates relevant low- and high-level features, resulting in an emergent category structure (Prince & Konkle, 2020). Accumulating research agrees on the inferior occipitotemporal regions as the potential neurobiological underpinnings of semantic categorisation of objects (Iordan et al., 2015), faces (Guntupalli et al., 2016) and places (R. A. Epstein & Baker, 2019). Other regions in the ITC involved in action observation and in representing ‘acting bodies’, including FFA, take part to scenes encoding (Groen et al., 2018). Accordingly,

Brooks et al. (2019) demonstrated that subjects' conceptual space predicts the neural pattern activation in the right FFA (Brooks, Chikazoe, Sadato, & Freeman, 2019). We may have observed stronger neural similarity between emotional categories in these regions, because of the influence of the precuneus, involved in valence representation and structurally connected with the ITC (Y.-H. Lin et al., 2020).

When we investigated the emotional and the neutral parts of participants' similarity space, we observed higher emotional similarity in the EVC, OPA and PPA, as well as in the dACC and anterior insula. OPA and PPA relate low-level visual features encoded in the EVC with the high-level aspects of the scene (R. A. Epstein & Baker, 2019) and may be modulated by regions that are sensitive to salience (anterior insula, dACC) (Lindquist et al., 2012), resulting in higher similarity. Interestingly, our finding that the insula represented the emotional, but not neutral, similarity space replicate those of Levine and colleagues who reported that it represented similarity ratings among emotional stimuli, although they were not controlled for semantic similarity and the ratings were of emotional rather than overall similarity (Levine et al., 2018). It would be worth exploring whether we would replicate the same results using Levine et al. (2018) instructions (Martina Riberto, Pobric, & Talmi, 2020).

Finally, the same effect was observed in the EVC, which relies on more fine-grained representations of the stimuli (Coutanche, Solomon, & Thompson-Schill, 2016), and encodes low-level visual features of the stimuli that afford the decoding of a broad range of emotions categories (Barrett & Bar, 2009). Specific combinations of low-level features (e.g., luminance) along with high-level information (e.g., presence of faces or scenes) can act as cues and afford specific categories of emotional response (Kragel et al., 2019). This might be paralleled by neural synchronization that connects the different neuronal populations involved in the processing of each feature (e.g., low and high level visual, emotional features) with the distant brain networks involved in each feature during the emotional experience (Sander, Grandjean, & Scherer, 2018).

We also expected that the orbitofrontal, ventral and dorsomedial prefrontal cortex were involved in representing the similarity space or just the emotional part. However, we did not find significant correlations with the behavioural data there, perhaps due to the implicit processing of affect in experiment 3. Nor did we

observe correlations with the amygdala, perhaps because it habituates quickly to repeated stimuli (Plichta et al., 2014).

No differences in perceived similarity between emotional and neutral pictures of realistic events.

In experiment 1, when thematic similarity was not controlled, we found higher similarity between emotional than neutral stimuli. In a valence-arousal space, emotional stimuli were placed closer to each other than neutral ones. The goodness of fit suggested that affective features were the most salient in similarity judgements. This result is in keeping with dimensional perspectives on emotions (Barrett & Russell, 1999) and recent empirical data (Cowen & Keltner, 2017), although our data cannot distinguish between effects based on valence or arousal dimensions. Strikingly, when we controlled for the higher thematic relatedness between emotional stimuli, by selecting stimuli from separate semantic categories, the rated similarity between stimuli from the two emotional categories was equivalent to that between those from the two neutral categories. Ratings clustered according to the four categories, and the goodness of fit dropped to fair, suggesting that the semantic meaning of each picture – not negative emotion - was the most relevant feature.

These findings accord with claims that participants' conceptual workspace comprises of integrated perceptual, affective, and semantic dimensions (Prince & Konkle, 2020). The evolved sensitivity to emotion, evident in the neural data, may be dampened when the context suggests it is less relevant, in keeping with previous literature attesting to the strong context effects on similarity (Goldstone et al., 1997). The relative contribution of semantic and affective features to overall similarity could be tested in future by collecting separate ratings of semantic or emotional similarity, or by manipulating the weight of the semantic and emotional dimensions. This opens up a new direction in semantic cognition research, which so far has not considered affective dimensions as key to semantic categorisation (Lambon Ralph, 2014).

Limitations

Our study presents several limitations that can be addressed in future works. First, we studied only negative emotions, and only two categories within each level of

affect. Second, stimuli were presented during a rapid event related design. Whilst a common approach, it might have influenced our results by increasing across-trial correlations (R. M. Visser et al., 2016), decreasing our statistical power. Finally, we cannot infer any causal role of emotions on neural similarity. Future studies could use TMS to further explore this aspect of the findings.

Conclusion

Stimuli that evoke negative feelings are perceived as more similar to each other unless care is taken to eliminate their taxonomic and thematic links. Once such semantic links are controlled, negative emotional and neutral stimuli may be judged as equally similar. A set of brain regions beyond those that are functionally specific to affect expressed emotional similarity preferentially. The stronger neural similarity between emotional pictures did not influence explicitly-perceived similarity in the same participants in the immediately-proceeding behavioural rating task, perhaps because the weights of the multiple dimensions of participants' conceptual workspace can change dynamically. Our findings may illuminate the clinically-relevant overgeneralisation bias in anxiety disorders. People with anxiety may have increased propensity to consider later, emotionally-similar, experiences as globally similar to the original fearful one, and thereby make maladaptive choices.

4.5 Supplementary Information

Emotional		**Neutral**	
205	016	104	165
127	038	100	035
226	075	150	095
238	007	066	249
235	022	146	057

Figure 4.2-1. Picture IDs from the NAPS database ('people' category), divided into emotional and neutrals (experiment 1).

Figure 4.3-1 corresponds to **Figure 2.1.**

5. Chapter: Increased neural similarity across sensory modalities after aversive conditioning.

Abstract

Generalisation, the extent to which a conditioned response is elicited in absence of an unconditioned stimulus, is an indirect measure of similarity. It increases with the resemblance among stimuli along perceptual or non-perceptual (e.g., semantic knowledge) dimensions. A better understanding of the neural mechanisms underlying generalisation might have clinical implication in anxiety disorders. After a traumatic experience, patients consider later experiences as similar to the original fearful one not because of their ostensible meaning, but their emotional similarity. Most of the studies pointed at overgeneralisation bias as a failure in perceptual discrimination between stimuli that predicted a negative outcome. However, it is still unclear whether aversive conditioning affect also the perception of similarity between threatening stimuli. A further debate concerned experience-dependent neural changes after aversive conditioning, and whether they generalised across sensory modalities. Multi-voxels pattern analyses (MVPA) showed a strengthening in the neural similarity structure in the occipitotemporal cortex of the stimuli that belonged to the same semantic category of CS+ after aversive conditioning. We hypothesised that these neural changes generalised from visual to auditory sensory modalities, and that they might underlie behavioural changes in similarity perception, resulting in higher similarity among threatening than not threatening stimuli. We tested these hypotheses in two aversive conditioning experiments, wherein participants rated which stimuli (i.e., images and vocalisation of mammals and birds) predicted money loss. In experiment 1, we focused on similarity perception, asking participants (n=20) to judge the similarity among visual stimuli before and after the conditioning. In experiment 2 (n=40), we explored neural similarity among brain activation patterns associated with threatening stimuli. We found increased neural similarity in the occipitotemporal cortex between visual threatening stimuli, and in the insula between visual and auditory threatening stimuli, despite no differences between threatening and non-threatening stimuli in similarity ratings after conditioning. Increased neural similarity might be the result of local and distance synchronisation processes between brain regions involved in fear generalisation.

This might be an efficient and functional mechanism, which may become maladaptive in individuals with anxiety disorders.

5.1 Introduction

Emotional similarity refers to the tendency to group stimuli together because they evoke the same feeling in us (Martina Riberto et al., 2019). It is a fundamental principle in cognition, as it supports core functions, such as categorisation (Barrett, 2017; Barsalou, 2017), memory and learning (Leal, Tighe, Jones, et al., 2014; Leal & Yassa, 2018; Talmi & McGarry, 2012). Research on emotional similarity may have clinical implications for the overgeneralisation bias in anxiety disorders. After a traumatic event, patients consider later experiences as similar to the original fearful one not because of their ostensible meaning, but their emotional similarity (Ahrens et al., 2016; Laufer et al., 2016). For example, we might judge two animals (e.g., cow and horse) similar, as they are both part of the semantic category 'mammals'. However, after an accident in a horse race, seeing (and perhaps hearing the sound of) a horse may trigger negative emotions (e.g., fear) and anxiety symptoms. Then, this reaction may be generalised to other exemplars of the same semantic category (e.g., cow), which became a threat too.

Generalisation, the extent to which we have a conditioned response (CR) in absence of an unconditioned stimulus (US), can be conceived as indirect measure of similarity, assuming a positive correlation between them. One way to investigate these constructs in an experimental setting is through aversive conditioning paradigms, wherein a previously neutral stimulus (e.g., horse) can elicit a CR (e.g., fear response) if that stimulus has been associated with an aversive US (e.g., accident at the horse race), becoming a conditioned stimulus (CS) (Fendt & Fanselow, 1999; Pavlov, 1927). After the conditioning, participants are shown with new generalisation stimuli (GS), which are not paired with the US, but might evoked the same CR, according to their similarity to the CS. A large body of studies (for a review see (Dymond, Dunsmoor, Vervliet, Roche, & Hermans, 2015)) explored generalisation as a function of similarity along perceptual dimensions. They pointed at overgeneralisation bias as a failure in perceptual discrimination between shapes (Lissek et al., 2009; Struyf, Zaman, Hermans, & Vervliet, 2017), or tones (Laufer & Paz, 2012; Resnik, Sobel, & Paz, 2011; Schechtman et al., 2010; Shalev, Paz, & Avidan, 2018) that predicted a negative (e.g., money loss, shock delivery, aversive pictures) than positive or neutral outcome. However, it is still unknown whether aversive conditioning also affect explicit similarity

judgements, resulting in higher similarity among threatening than not threatening stimuli.

Yet, other aspects affect fear generalisation, such as emotional intensity and semantic knowledge. Dunsmoor et al. (2011) reported higher skin conductance response (SCR), a measure of physiological arousal, and wider generalisation among morphed faces, similar to the one paired with an electric shock (CS+), but only with high emotional intensity (Dunsmoor, Prince, et al., 2011). Others explored the role of semantic knowledge on fear generalisation, showing higher accuracy in memory recognition (Dunsmoor, Martin, & LaBar, 2012), and wider generalisation of episodic memory (Starita et al., 2019) for GS from the CS+ category than the CS- category. This suggest that threat learning improved memory recognition and promotes generalisation of episodic memory in healthy controls. It is an adaptive mechanism, functional to survival, which may become maladaptive when old threat memories are inappropriately reactivated in secure situations, for example in individuals with high trait anxiety.

Studies in healthy controls showed a positive correlation between trait anxiety scores and memory for the CS− stimuli (Cohen et al., 2019), and with the SCR to the preconditioned stimulus (i.e., image of a spider) that was semantically related to the CS+ (i.e., image of a web) (Dunsmoor, White, & LaBar, 2011). In addition, others found a significant difference in valence and arousal ratings following conditioning between patients with social anxiety and controls, with lower valence and higher arousal ratings provided by the former (Ahrens et al., 2016). However, no studies explored whether trait anxiety amplifies the emotional impact of the GSs after conditioning in healthy controls.

Different mechanisms have been proposed as neurobiological underpinnings of fear generalisation, including pattern separation and pattern completion occurring in the hippocampus (Kheirbek, Klemenhagen, Sahay, & Hen, 2012; Lange et al., 2017; Lissek et al., 2014). Following fear conditioning, when exposed to a GS (in the previous example, a cow) that resemble the CS+ (e.g., a horse), the overlap between patterns of brain activity representing GS and the previously encoded CS+ is assessed through hippocampal based ‘schematic matching’. Given sufficient overlap, CA3 neurons in the hippocampus induce pattern completion, whereby a subset of cues from a previous experience (i.e. CS+) activates the stored pattern representing that experience. In case of insufficient overlap

between neural representations of GS and the CS+, the dentate gyrus neurons initiate pattern separation, which takes similar patterns of neural activity and converts them into distinct representations (Fenton, 2007). These complementary computations are carried out in dynamic networks that involved different brain regions other than the hippocampus, including the thalamus, brain structures associated with fear excitation (e.g., the amygdala, anterior insula, dACC) and fear inhibition (e.g., MPFC) (for a review see (Dunsmoor & Paz, 2015; Fullana et al., 2016; Gross & Canteras, 2012)).

More recently, multi-voxels pattern analyses (MVPA) of fear-conditioning have shown that the neural representations of fear-conditioned stimuli that belong to the same semantic category (i.e., 'animals'; 'objects') become more similar to each other in the occipitotemporal cortex (Dunsmoor et al., 2013), in the amygdala (R. M. Visser et al., 2013) and the superior frontal gyrus (R. M. Visser et al., 2011), and more similar to the US in the insula (Onat & Büchel, 2015). Others reported increased neural similarity between stimuli associated with reward (Zeithamova et al., 2018) and pain (Wagner et al., 2020) in the hippocampus. This survival mechanism has been explored only using visual stimuli.

However, an open research question concerns whether fear generalize across (visual and auditory) sensory modalities (similarity across sensory modalities), resulting in higher neural similarity between GSs that predict the same outcome of the CS+, regardless of their sensory modality. In order to answer our research questions, in experiment 1 we explored whether fear conditioning biased explicit similarity judgements, with higher similarity between threatening than not threatening stimuli. In experiment 2, we investigated whether this behavioural change is paralleled by increased neural similarity between exemplars in GS+ than GS- across sensory modalities following aversive conditioning in healthy controls.

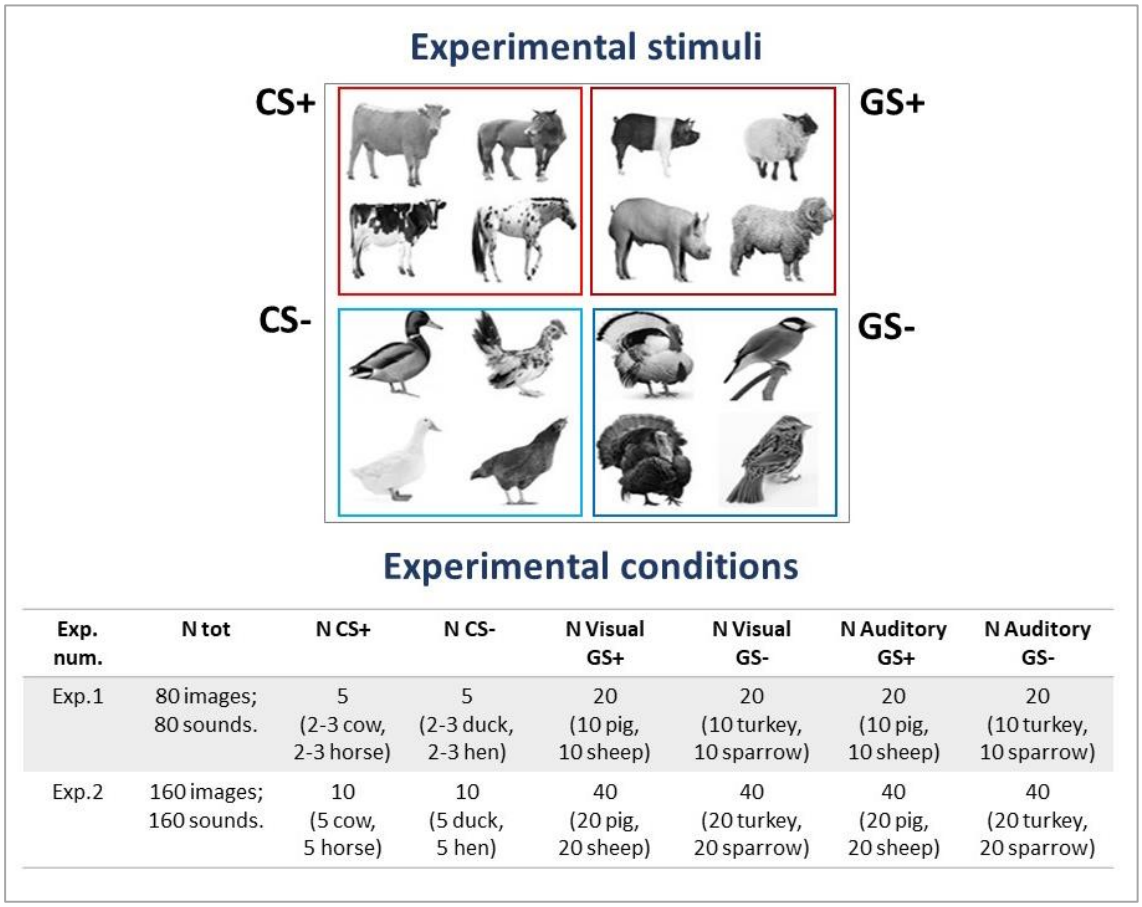
5.2 Materials and Methods

Participants

A total of 60 right-handed participants were recruited from the Weizmann Institute of Science (Israel) to take part in the study (female: 30; age range, 21 –58 years; mean age, 30.30 years; SD, 7.34) (Experiment 1: 20 participants, 14 females; Experiment 2: 40 participants, 16 females). All participants had normal or corrected-to-normal vision, and were older than 18 years. They gave informed consent prior to the experiment and have been reimbursed for their participation. The exclusion criteria were: a history of neurological (e.g., head injury or concussion) or psychiatric conditions (e.g., depression, anxiety), drug or alcohol abuse, or regular medication that could influence emotional processing. The study was approved by the ethics board of the Weizmann Institute of Science (protocol number 0287–09-TLV).

Materials

Experiment 1 comprised of 80 images and 80 sounds of animals (‘mammals’: 40 images, 40 sounds; ‘birds’: 40 images, 40 sounds). These were grouped into 8 basic-level categories (i.e., mammals: cow, horse, pig, sheep; birds: duck, hen, turkey, and sparrow). Four of them served as conditioning stimuli (CS) (e.g., mammals CS+: cow, horse; birds CS-: duck, hen), and the remaining four categories as generalisation stimuli (GS) (e.g., mammals GS+: pig, sheep; birds GS-: turkey, sparrow). Each basic-level category consisted of 10 images and 10 sounds of different exemplars (e.g., different breeds). As Experiment 2 involved an fMRI task that required a rich conditions design (Nili et al., 2014), we doubled the number of both images and sounds. Specifically, we selected 160 images and 160 sounds of animals (‘mammals’: 80 images, 80 sounds; ‘birds’: 80 images, 80 sounds), and each basic-level category consisted of 20 images and 20 sounds of different exemplars, as shown in Figure 5.1.



Experimental conditions

Exp. num.	N tot	N CS+	N CS-	N Visual GS+	N Visual GS-	N Auditory GS+	N Auditory GS-
Exp.1	80 images; 80 sounds.	5 (2-3 cow, 2-3 horse)	5 (2-3 duck, 2-3 hen)	20 (10 pig, 10 sheep)	20 (10 turkey, 10 sparrow)	20 (10 pig, 10 sheep)	20 (10 turkey, 10 sparrow)
Exp.2	160 images; 160 sounds.	10 (5 cow, 5 horse)	10 (5 duck, 5 hen)	40 (20 pig, 20 sheep)	40 (20 turkey, 20 sparrow)	40 (20 pig, 20 sheep)	40 (20 turkey, 20 sparrow)

Figure 5.1. Top: experimental stimuli for Experiment 1-2. We selected images and sounds of animals from two superordinate categories ('mammals and 'birds'), and divided them into 8 basic-level categories: cow, horse, pig, sheep, duck, hen, turkey and sparrow. For each participant, two basic-level categories within each superordinate category served as CS and two as GS. In this case, images of cow and horse served as CS+ (light red square), and images and sounds of pig and sheep as GS+ (dark red square); images of duck and hen represented the CS- (light blue square), and images and sounds of turkey and sparrow as GS- (dark blue square). Bottom: Number of trials for each experimental condition, divided into experiment 1 and 2, averaged across sessions. The number of stimuli in experiment 2 (n=160 within each sensory modality) is doubled than those in experiment 1 (n=80 within each sensory modality). Abbreviations: CS, conditioned stimuli; GS, generalisation stimuli.

In both experiments each stimulus was presented only once. We selected the visual stimuli using Google images, the sounds from www.soundsnap.com, www.epidemicsound.com, and from publicly available resources on the internet. Images were grey-scaled and presented on a white background (300 x 200) for 3 seconds; sounds duration was shortened to 3 seconds and noise background was removed using Audacity software (www.audacityteam.org). Stimuli were presented using Psychtoolbox-3 (http://psychtoolbox.org). The loss of 1 and 2 Israeli shekels (NIS) served as US in Experiment 1 and 2, respectively.

Procedure

A graphical representation of the general experimental procedure is shown in Figure 5.2. Participants performed an aversive conditioning task, wherein they

learned the association between CS (images only) and US (money loss), and generalise it to new auditory and visual stimuli (GS). In case of loss, the money were taken from an initial amount (Experiment 1: 50 NIS; Experiment 2: 250 NIS) that they received before starting the experiment. The allocation of animal to CS condition was counterbalanced across participants. Experiments 1-2 differed in few aspects. In Experiment 1 only, participants completed a multi-arrangement (MA) task to judge the similarity between images before and after the conditioning task. In Experiment 2, participants filled the STAI-T and STAI-S questionnaire (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983), before the aversive conditioning fMRI task, wherein their pupil dilation was recorded through the eye tracking, as neurophysiological measure of aversive conditioning. After the scan, they performed a surprise valence and arousal rating task of some of the images. In the end of the experiment, we reimbursed participants for their time (Experiment 1: 40 NIS; Experiment 2: 200 NIS), and explained them the real purpose of the study.

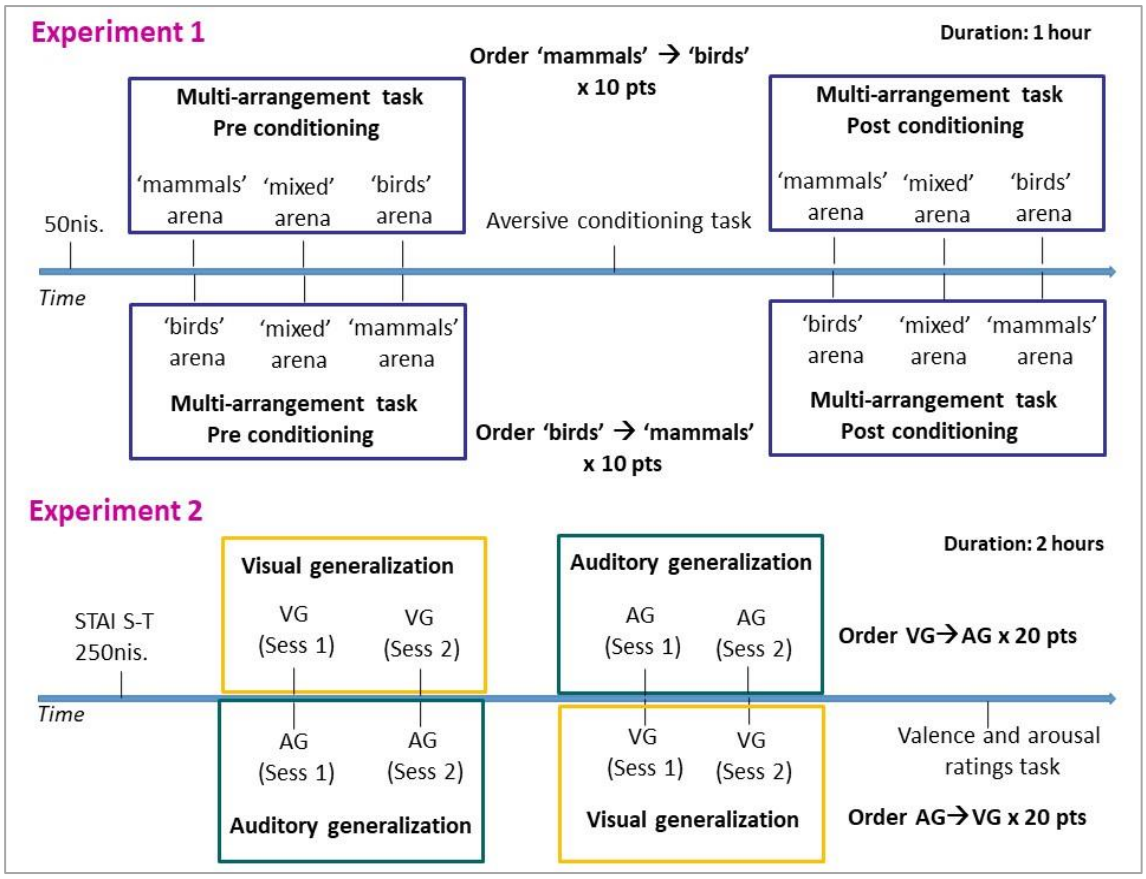


Figure 5.2. Top: general procedure in experiment 1. After receiving 50 Israeli shekels (NIS) and being instructed for the task, 20 participants judged the similarity of images of animals by performing the multi-arrangements task before and after aversive conditioning task. They dragged and dropped the images into

circular arenas, one for each condition ('mammals, 'birds', and 'mixed' with 'mammals' and 'birds' in it), wherein the proximities reflected the similarity among images. The order of the arenas was counterbalances across subjects, such that half judged first the 'mammals' and then the 'birds', and half the opposite. The order for each participant was the same before and after the conditioning. In between, participants performed an aversive conditioning task, which has the same structure of the fMRI task in experiment 2. The entire experiment lasted approximately one hour. Bottom: general procedure in experiment 2. We instructed 40 participants about the fMRI aversive conditioning task, after giving them 250 NIS and asking to fill the STAI_S and STAI_T questionnaire. During the scan, they learned the association between images (VG) or sounds (AG) and money loss vs saving. The aversive conditioning task was divided into 4 sessions (two VG and two AG), and their order was counterbalanced across participants, such that half performed first VG and then AG, and half the opposite. After the MRI, participants performed a surprise valence and arousal rating task outside the scan. The entire experiment lasted approximately 2 hours. Abbreviations: CS, conditioned stimuli; GS, generalisation stimuli.

Multi-arrangements task

In Experiment 1, participants judged the similarity among the pictures by using the multi-arrangements (MA) task, a quick and efficient task for acquiring similarity judgements in experiments with a relatively large number of stimuli. According to Kriegeskorte and Mur (2012), the MA has high test-retest reliability (r=0.81) as well as external validity (Kriegeskorte & Mur, 2012). Because of time constraints, images were divided into 'even' and 'odd', for a total of 40 images within each subset (20 mammals and 20 birds). Each participant was randomly assigned to one of the subsets (10 participants for each subset). In each trial of the MA task, 14 stimuli were presented along the perimeter of a circle, or 'arena', on a computer screen. In order for participants to focus on the similarity both within and between basic-level categories, we opted for splitting the images into 3 different arenas, one for mammals and the other for birds, each with 5 exemplars for each basic-level category, and one 'mixed' arena with 4 exemplars of mammals and 4 exemplars of birds, selected randomly from each of the 8 basic-level categories. The purpose of the mixed arena was to compare within-category and between-category similarity. Participants had unlimited time to drag and drop the stimuli in each arena according to their similarity, such that similar stimuli were placed close to each other and dissimilar stimuli apart. We counterbalanced the order of the arenas, which remained the same before and after conditioning, such that half of the participants arranged first the 'mammals' and then the 'birds' arenas, and half the opposite. The 'mixed' arena was always in between the others. We instructed participants to focus on any aspect they considered relevant for the judgements. A trial ended when participants arranged all the stimuli in the arena. Subsequent trials started with another subset of stimuli to be arranged, selected by using the 'Lift-the-weakest algorithm for adaptive design of item subsets'. This method optimises trial efficiency by adaptively selecting item subsets whose dissimilarity

estimates presented the weakest evidence. The task ended after approximately 20 minutes, when participants judged all the possible combinations among stimuli within each arena.

Valence and arousal rating task

In Experiment 2, after the conditioning, participants completed a surprise valence and arousal rating task outside the scan, following the procedure suggested by Lang et al. (2008) (Lang et al., 2008). This task involved 10 randomly selected images for each basic-level category, presented in a pseudorandom order (total= 80 images: 40 'mammals' and 40 'birds'). Each trial started with a central fixation cross for 500 ms. Then, participants viewed one of the images presented in the centre of the screen, and rated each pictures on two 9-points scale (valence scale: 1, negative emotions; 9, positive emotions; 5 neutrals. Arousal scale: 1, relaxed; 9, aroused; 5 neutral). We instructed participants to respond as quickly as possible by clicking the appropriate number key, and informed them that there was not a right or wrong answer.

Aversive conditioning task

As shown in Figure 5.2, participants took part in 4 sessions, each of which lasted approximately 10 minutes. Each session included 10 Pavlovian conditioning (PC, always visual) and 40 generalisation trials (either visual, or auditory). During PC trials, 2-3 visual CS for each basic-level category were displayed on a blank screen for 3 seconds. While the stimulus was presented, participants rated whether it predicts either the loss or the saving of 2 shekels (i.e., lose vs saving) and their degree of confidence (i.e., sure vs almost sure), by pressing one of the four keys on the response box. Then, according to the nature of the CS, a feedback appeared on the screen (CS+: '*you lost 2 NIS*'; CS- '*you saved 2 NIS*'). A fixation cross-followed the offset of each trial. Generalisation trials were identical to PC trials, except that the US was not shown. While participants were conditioned using visual stimuli, generalisation trials included either visual or auditory stimuli.

Two of the sessions involved visual generalisation (VG1 and VG2), while the other two involved auditory generalisation (AG1 and AG2). VG1 and AG1 started with 10 PC trials (5 CS+ and 5 CS-), followed by 40 generalisation trials (20 GS+ and 20 GS-). In VG2 and AG2, PC and generalisation trials were intermixed. Stimuli were presented in a pseudorandomised order, with no more than 2 consecutive

items from the same category occurring in a row. The purpose of PC trials was to enable learning, but otherwise, they were not the focus of our main hypotheses, and they were therefore fewer in number than GS trials. We counterbalanced the order of VG and AG across participants, such that 20 participants performed first VG and then AG, and 20 the opposite (first AG, and then VG). We also counterbalanced the allocation of category (birds or mammals) to the CS+/GS+ and CS-/GS2 conditions.

We instructed participants that their goal was to understand which images and sounds of animals predicted the loss vs the saving of 2 NIS. In case of loss, the money were taken from the initial amount of 250 that they received before the scan. To increase engagement, participants were told that pressing the wrong key will double their loss. Participants received a mean estimation of their loss at the end of the second and of the fourth sessions. This estimation was purposefully false, and was included to increase the emotional impact of loss on participants. In the end of the task, they were reimbursed the same amount for their time, regardless of their performance during the task.

The aversive conditioning task used in experiments 1- 2 were very similar, apart for the length of the inter-stimulus interval (ISI) (Experiment 1: 3 seconds; Experiment 2: 6-7 seconds), the number of stimuli (Experiment 1: 80 images and 80 sounds; Experiment 2: 160 images and 160 sounds) and the size of the US (Experiment 1: 1 NIS; Experiment 2: 2 NIS).

MRI data acquisition

MRI procedure

Images were acquired on whole body 3T MRI scanner (Trio TIM, Siemens, Germany) with a 12-channel head coil. Functional images were acquired with a susceptibility weighted EPI sequence (TR/TE=2000/30 ms, flip angle=75 degrees, voxel dimensions=3x3x3.5 mm, 192 slices) in 4 separate scanning sessions (up to two minutes between sessions). Anatomical T1-weighted images were acquired after the functional scans (MPRAGE, Repetition time (TR)/Inversion delay time (TI)/Echo time (TE)=2500/900/2.32 ms, flip angle=8 degrees, voxel dimensions=1 mm isotropic, 32 slices).

Eye Tracking

The pupil diameter was acquired with an EyeLink 1000+ eye tracker video-based system (SR Research). The camera was positioned just below the monitor in the MRI scanner, recording binocularly pupil size (in arbitrary units) with a sampling frequency of 1000 Hz (Alamia, VanRullen, Pasqualotto, Mouraux, & Zenon, 2019). The standard five-point calibration procedure for the EyeLink system was conducted prior to the first session (Liao, Yoneya, Kidani, Kashino, & Furukawa, 2016). As pupil measurements depend on the gaze angle (Hayes & Petrov, 2016), participants were asked to fixate either the stimulus or a fixation cross positioned in the centre of the screen in all experiments. No explicit free viewing or blinking periods were included in the tasks.

Data analysis

In Experiment 1, we expected higher similarity (lower dissimilarity) within than between categories. This prediction serves as manipulation check, since a good category boundary simultaneously maximize the within-category similarity, and minimize the between categories similarity. Our main hypothesis was that aversive conditioning increased the similarity among threatening stimuli, as showed at the top of Figure 5.3. This hypothesis applied also to the neural data in Experiment 2. Specifically, we predicted higher neural similarity in GS+ than GS- regardless of their sensory modality, as showed at the bottom of Figure 5.3. We tested this effect within visual GSs to replicate previous findings. In addition, for the first time we explored it across sensory modalities. We tested this effect within auditory GSs, and between visual and auditory GSs. In the former analysis, we investigated whether auditory threatening GS become more similar to each other, in the latter whether auditory and visual threatening GS come to be alike, following aversive conditioning. The last analysis allowed us as manipulation check to test higher similarity between the image and the sound of a mammal (i.e., GS+) than between the image of a mammal and the sound of a bird (i.e., vGS+/aGS-), or between the image of a bird and the sound of a mammal (i.e., (vGS-/aGS+). The same was valid for the GS-, as showed at the bottom right of Figure 5.3. As additional manipulation check, we expected that these effects were paralleled by differences in pupil diameter (PD), BOLD signal change, and valence and arousal ratings. We predicted higher recruitment (higher activation) of brain regions involved in semantic processing and aversive conditioning associated with threatening than

not- threatening stimuli, wider pupil diameter as well as lower valence and higher arousal ratings. Additional details about the statistical analyses are reported in the following sections.

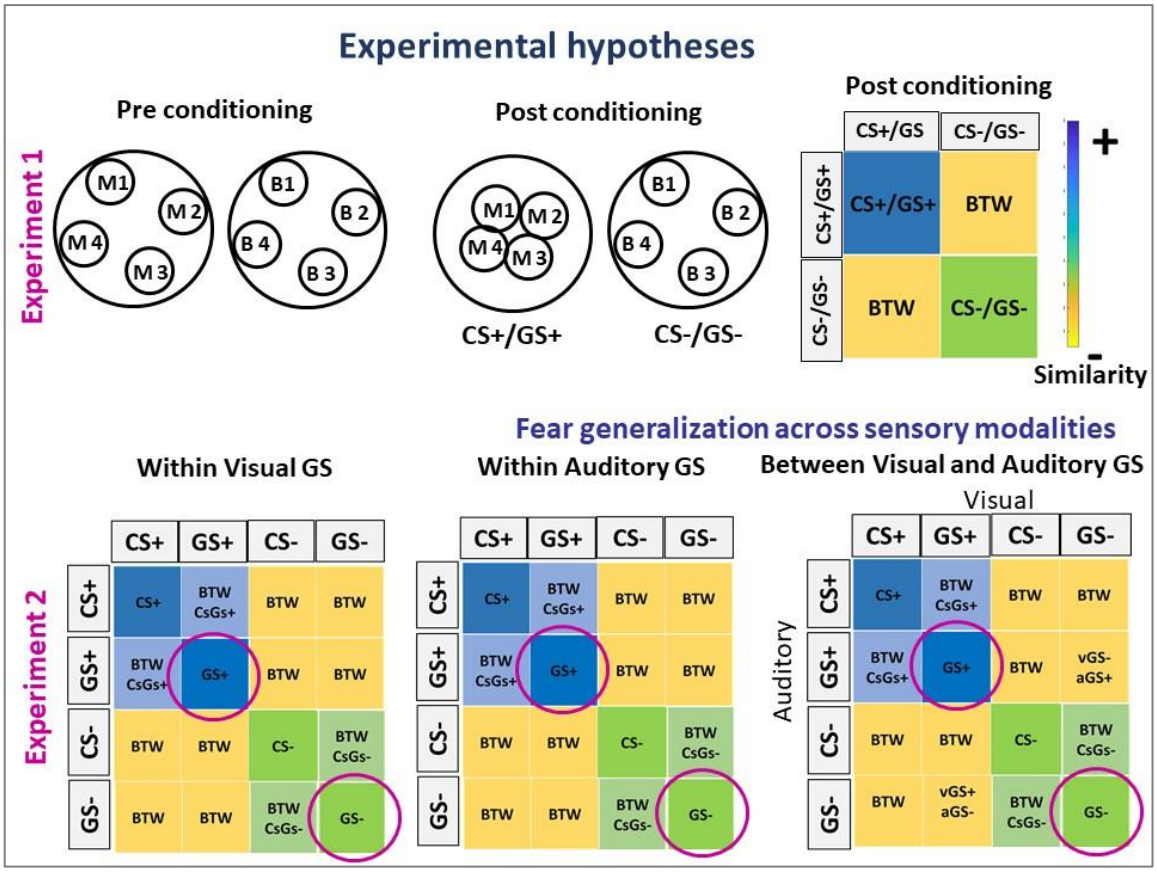


Figure 5.3. Top. In experiment 1, we expected higher similarity within than between categories, and that aversive conditioning would increase the similarity in CS+/GS+ than CS-/GS-. The similarity in CS+/GS+ was calculated as Euclidean distance between items from CS+/GS+ category in the 'CS+/GS+' arena (in this case, the 'mammals'), pre and post conditioning. The similarity in CS-/GS- was calculated as Euclidean distance between items from CS-/GS- in the 'CS-/GS-' arena (in this case, the 'birds'), pre and post conditioning. The similarity between CS+/GS+ and CS-/GS- was calculated as Euclidean distance between 'CS+/GS+' and 'CS-/GS-' in the 'mixed' arena. Blue colour denote high similarity (low Euclidean distance), and yellow low similarity (high Euclidean distance). Bottom. In experiment 2, we predicted higher neural similarity in GS+ than GS- (denoted with pink circles), within visual modality and across visual and auditory sensory modalities. In the RDM within visual GS and within auditory GS, the similarity in GS+ and GS- was measured as correlational distance (1- Spearman's correlation) in GS+ and in GS-, respectively, averaged across sessions, in VG and AG separately. These conditions were also valid in the RDM between visual and auditory GSs (e.g., GS+ represented the similarity between visual and auditory GS+). In the latter matrix, we also tested as manipulation check higher similarity in GS+ than in vGS+/aGS- and vGS-/aGS+. The same was valid for GS- . vGS+/aGS- and vGS-/aGS+ were measured as correlational distance (1- Spearman's correlation) between visual GS+ and auditory GS-, and between visual GS- and auditory GS+, averaged across sessions. Abbreviations: M, Mammals; B, Birds; CS, Conditioned stimuli; GS, generalisation stimuli; BTW, Between CS+ and CS-/ Between GS+ and GS-; CS (+,-) within Conditioned stimuli; GS (+,-) within Generalisation stimuli; BTW CsGs (+,-), between CSs and GSs; vGS+/aGS-, between visual GS+ and auditory GS-; vGS-/aGS+, between visual GS- and auditory GS+.

Behavioural and neurophysiological data.

Behavioural and neurophysiological data were analysed using SPSS (IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp) and Matlab R2018a (MATLAB 2018a, The MathWorks, Inc., Natick, Massachusetts, United States).

Analysis of similarity (MA task).

In Experiment 1, similarity was measured as Euclidean distance between stimuli in the arena. Specifically, in the first two arenas (one for 'mammals' and one for 'birds'), a *partial* RDM is estimated at the end of each trial, showing the Euclidean distance between stimuli within each trial. After participants judged the similarity among all the stimuli in each arena, two *global* 20 x 20 RDMs (one for 'mammals' and one for 'birds') were estimated by averaging the partial RDMs with an iterative rescaling. This scaling procedure takes into account that in each trial participants focused on a specific subset, and that, therefore, there is not a permanent relationship between screen distance and dissimilarities across trials (see (Kriegeskorte & Mur, 2012) for details). The *global* 8 x 8 RDM from the 'mixed' arena (with both mammals and birds) is estimated after one trial, as participants judged all the possible combinations in one single trial. Then, we extracted from each participant's *globa*l 20 x 20 RDMs the mean and the standard deviation of the conditions of interest: dissimilarity within 'mammals', dissimilarity between subordinate-level categories of 'mammals' (e.g., between cow and horse), dissimilarity within 'birds' and dissimilarity between subordinate-level categories of 'birds' (e.g., between duck and hen), both pre and post aversive conditioning. We also estimated in each participant's *globa*l 8 x 8 RDM the dissimilarity within 'mammals', dissimilarity within 'birds' and dissimilarity between 'mammals' and 'birds', both pre and post aversive conditioning. These were used as dependent variables in 2 repeated- measures ANOVAs. In the former ANOVA, we entered as dependent variables the mean of the conditions extracted from the 8 x 8 RDM (i.e., dissimilarity within 'birds', within 'mammals', and between 'birds' and 'mammals'), and we tested the manipulation of similarity, specifically, the hypothesis that dissimilarity within categories will be lower than between categories. The latter (2 x 2 ANOVA: time [pre vs post] x stimulus type [+ vs -]) was used to investigate the main hypothesis, that is, higher similarity (lower dissimilarity) among threatening stimuli (CS+ and GS+) after conditioning. We tested this hypothesis using the dissimilarity estimates from the *globa*l 20 x 20 RDMs.

Analysis of aversive conditioning task.

We measured aversive learning using the accuracy in ratings of loss expectancy. Accuracy scores were calculated by dividing the number of correct answers, that

is, when participants rated they were either sure or almost sure to lose money in a CS+/GS+ trials (and vice versa for CS-/ GS-), by the number of stimuli within each condition. Accuracy was measured separately for PC, VG, and AG trials, and averaged across sessions (4 sessions in PC trials, 2 sessions for VG and 2 sessions for AG). Successful learning was defined as an accuracy score higher than chance level (50%) across CSs during PC trials. In a similar manner, we defined successful generalisation as above-chance accuracy, computed separately for the VG and AG condition. It is worth noticing that participants' responses are considered 'correct' in the eyes of the researcher, though they are not objectively 'correct'. We analysed learning and generalisation with three one sample t-tests (separately for PC, VG and AG), testing the null hypothesis of no difference between the average accuracy and the chance level (50%). We also expected increased accuracy scores and decreased RTs over time in each of these three conditions. We tested this hypothesis by entering average accuracy and latency within the first and the second sessions in three repeated-measures ANOVAs (separately for PC, VG and AG, one for accuracy and one for latency), with time as within-subject factors.

Analysis of valence and arousal ratings.

We estimated average valence and arousal ratings for CS+, CS-, GS+ and GS-. These were entered as dependent variables in paired t-tests (one for the CSs and for the GSs), separately for valence and arousal. We also estimated the Spearman's correlations between valence GS+ and trait anxiety scores (STAI_T), expecting a negative correlation, as well as between arousal GS+ and STAI_T, predicting positive correlation.

Analysis of Pupil diameter (PD)

We measured the PD as manipulation check that the aversive conditioning occurred. We analysed only right-eye data, since data from both eyes showed a similar pattern. Time series were analysed from the beginning of the first stimulus presentation until the last trial. Data acquired during blinks were treated as missing. To obtain a mean response, we extracted data segments following trial onset, response time, and US time in PC trials only, and averaged these segments separately first within and then across participants. The first segment (i.e., onset) was maximum 3 seconds long; the length of the second one (i.e., response time)

depended on participants' response time, thus ranging from 1 second to few milliseconds; the last one (i.e., US time) lasted 1.5 seconds. These were z-scored transformed (by subtracting the mean and dividing by the standard deviation) within each session to account for between-subjects variance in overall pupil size (Korn, Staib, Tzovara, Castegnetti, & Bach, 2017). We expected to observe wider PD for threatening stimuli (i.e., CS+ and GS+) than not threatening stimuli (i.e., CS- and GS-).

We predicted that this effect might follow trial onset, response time, and US time. In order to test this, we estimated average PD for our conditions of interest, separately for each sensory modality, that is, PD GS+ (onset), PD GS+ (response), PD GS- (onset), and PD GS- (response), averaged across sessions. The same conditions were valid in PC trials, with the addition of PD when US was delivered (i.e., PD CS+ (US) and PD CS- (US)). We entered them as dependent variables in different paired t-tests with stimulus type as within-subject factor, one for each segment (separately for PC, VG, AG).

Neuroimaging data analysis

Neuroimaging data were pre-processed and analysed using Statistical Parametric Mapping (SPM12) (http://store.elsevier.com/product.jsp?isbn=9780123725608) and MATLAB R2018a (MATLAB 2018a, The MathWorks, Inc., Natick, Massachusetts, United States). Functional images were slice-time corrected to reduce the mismatching between acquisition timing of different slices, and realigned to a reference (mean) image to minimize the variance due to head movements. These were then coregistered to the high-resolution T1-weighted structural image, which was coregistered and normalized to MNI space. Finally, functional images were normalized to a standard template volume based on the Montreal Neurological Institute (MNI) reference brain to achieve a more precise comparison across individuals. Spatial smoothing was performed only on functional data analysed with a conventional univariate approach using a 6-mm full-width at half-maximum isotropic Gaussian kernel. No spatial smoothing was carried over on the multivariate functional data, according to the standard practices for MVPA studies (Haxby et al., 2001; Kriegeskorte, Mur, et al., 2008b).

Individual-level model for RSA analysis

After preprocessing, functional data from each voxel were analysed using the general linear model (GLM). We modelled one GLM for each sensory modalities (visual GLM and auditory GLM), which comprised two separate sessions (i.e., visual GLM: VG1 and VG2; auditory GLM: AG1 and AG2). Each GLM comprised of 100 stimuli (50 for each session). Each stimulus was modelled as a separate event beginning with picture or audio presentation onset, using the canonical function in SPM12, and included in the model as regressor of interest (100 regressors for each GLM: 10 CSs and 40 GSs). Six motion correction parameters were also modelled within each session, and included in the model as regressor of no interest. From this GLM analysis, we obtained a single beta image for each stimulus. Contrast images for each stimulus against the implicit baseline were generated based on the fitted responses, and averaged across sessions. The resulting 100 T-contrast images within each sensory modality were used as inputs for RSA.

Individual-level models for univariate analyses

Although our hypotheses were specific to the multivariate representations, we also performed conventional univariate analyses as manipulation check. We tested whether our study replicated previous findings showing higher recruitment associated with threatening than not-threatening stimuli in brain regions involved in semantic processing and fear excitation, and lower recruitment in regions associated with fear inhibition. Individual-level models for univariate analyses were almost identical to those used for the RSA, the only difference being that instead of modelling 100 stimuli within each sensory modalities (visual GLM and auditory GLM), here each condition (i.e., CS+ and GS+, CS- and GS-) was modelled as separate condition (4 regressors per session) beginning with each picture/sound presentation onset, using the canonical function in SPM12. We included Reaction Times (RTs) as parametric modulator for each regressor, to take into account differences in latency between conditions. In the contrast images, we focused only on GSs stimuli, as they are of interest in this study. Contrast images for GS+ than GS- conditions against the implicit baseline were generated based on the fitted responses, and averaged across sessions, within each GLM separately. The resulting two T-contrast images (i.e., GS+ and GS-) within each sensory modality were used as inputs for univariate group analyses.

ROIs definition.

We defined the regions of interest (ROIs) by using the Automated Anatomical Labelling (AAL) template in WFU Pickatlas toolbox (https://www.nitrc.org/projects/wfu_pickatlas), Anatomy toolbox (https://www.fil.ion.ucl.ac.uk/spm/ext/#AAL), and Stanford functional atlas (https://findlab.stanford.edu/functional_ROIs.html). We used WFU_Pickatlas toolbox to define the Dorsomedial Prefrontal cortex (DMPFC) as Brodmann area 8 and 9, and the AAL template in WFU_Pickatlas for the bilateral Inferior Temporal cortex (ITC), the temporal pole (TP), the Hippocampus (HIP), the dorsal and ventral striatum, the Anterior and Middle Cingulate cortex (ACC, MidCC), the Thalamus, the Insula, the bilateral Orbitofrontal cortex (superior, middle, inferior and medial OFC), the Superior Frontal gyrus (SFG), and the Middle Frontal gyrus (MiFG). The bilateral basolateral Amygdala was defined using Anatomy toolbox, and the bilateral high-level auditory cortex with Standford functional atlas. Finally, the bilateral Inferior Occipital gyrus (IO) and the Fusiform gyrus (FG) as an 8 mm sphere around the coordinates reported by Dunsmoor et al. (2013) (Dunsmoor et al., 2013). We combined all these ROIs into a single ROIs mask to use as inclusive mask in the univariate group analyses. In addition, we used each ROI separately to investigate the main hypothesis (higher neural similarity within GS+) across sensory modalities, except for the ITC, FG and IO, wherein we tested the main hypothesis within visual modality, and the high-level auditory cortex within auditory modality.

Univariate group analyses

The individual level contrasted images were entered as input in full-factorial designs, one for visual and one for auditory modality, with group as between subject factors (i.e., whether participants performed first VG or AG) and condition as within subject factors (i.e., GS+ vs GS). We then compared GS+ and GS- in both directions (T1: GS+ > GS-; T2: GS+ < GS-) in the inclusive ROIs mask.

RSA group analyses: quantifying neural similarity across sensory modalities

We predicted higher neural similarity (lower dissimilarity) in GS+ than GS- across sensory modalities, as showed at the bottom of Figure 3. We first explored the

increased in neural similarity in visual GSs to replicate previous findings in all the mentioned ROIs, apart for the high-level auditory cortex that is modality specific. For each subject and ROI, we computed a brain activation-pattern RDM, where each entry represented the correlational distance (1- Spearman's correlation) between brain activations across voxels in each ROI, and the rows and the columns represented the visual experimental stimuli. We will refer to this as visual ROI RDM. It is symmetric about a diagonal of zeros, and resulted in 4950 cells in the lower triangular part that reflected the pairwise dissimilarity of the response patterns associated with the stimuli for each ROI. Then, within each participant ROI RDM, we calculated the mean and the standard deviation of the conditions of interest, including dissimilarity in GS+ (GS+) and in GS- (GS-). We entered them as dependent variables in paired t- tests, one for each ROI, testing the null hypothesis of no difference between GS+ and GS-.

Then, we investigated higher neural similarity in GS+ than GS- across sensory modalities. We tested this between auditory GSs, expecting higher similarity in auditory GS+ than GS-. We computed the auditory ROI RDM in the same way as the visual ROI RDM, except for the rows and the columns that represented the experimental stimuli in auditory modality. This effect was tested in all the mentioned ROI, apart for the ventral-visual stream that carried out modality-specific computations. In addition, we also tested fear generalization between visual and auditory GSs in all the ROIs with the exceptions of ventral-visual stream and high-level auditory cortex. To compute the RDM between visual and auditory GSs, for each participant we estimated the correlational distance between visual and auditory ROI RDMs (e.g., visual Amygdala RDM and auditory Amygdala RDM). This resulted in a 100 x 100 matrix, wherein each entry represented the correlational distance between brain activations across voxels associated with each stimulus in visual and auditory modality (within each ROI), the rows represented the experimental stimuli in auditory modality, and the columns in visual modality. From this RDM, we extracted the previously mentioned conditions of interest (GS+, GS-), the dissimilarity between visual GS+ and auditory GS- (vGS+/aGS-), and the dissimilarity between visual GS- and auditory GS+ (vGS-/aGS+), as showed at the bottom of Figure 3. As manipulation check, we expected higher similarity (lower dissimilarity) in GS+ than vGS+/aGS-and vGS-/aGS+. The same was valid for GS-. In other words, we predicted higher similarity between the

image and the sound of a mammal (i.e., GS+) than between the image of a mammal and the sound of a bird (i.e., vGS+/aGS-), or between the image of a bird and the sound of a mammal (i.e., (vGS-/aGS+)

We entered the conditions of interest as dependent variables in paired t- tests, one for each ROI and comparison of interest. The RSA was performed using the MRC-CBU RSA toolbox for MATLAB (http://www.mrc-cbu.cam.ac.uk/methods-and-resources/toolboxes).

5.3 Results

Behavioural and neurophysiological data

Accuracy and RTs

In experiments 1-2 aversive learning was evaluated using the accuracy in ratings of loss expectancy, and PD in experiment 2 only. In experiment 1, on average participants learned the association between US and CS, evident in higher accuracy scores than chance level across sessions ($t(19) = 7.94$; $p<0.001$, $d= 1.78$). In addition, accuracy scores improved over time, with higher accuracy in the last than in the first five trials ($t(19) = -2.85$; $p=0.010$, $d= 0.65$). However, we did not observe significant decrease in RTs over time. Successful visual (VG) and auditory (AG) fear generalization to new exemplars was measured as in the PC trials. On average, participants generalised fear to new exemplars, evident in higher accuracy scores than chance level, in both VG ($t(19) = 10.61$; $p<0.001$, $d= 2.35$) and AG ($t(19) = 20.01$; $p<0.001$, $d= 4.3$). Accuracy increased over time in AG only ($t(19) = -3.03$; $p=0.007$, $d= 0.45$). As in PC trials, RTs did not change over time.

In experiment 2, participants learned the association between US and CS, evident in higher accuracy scores than chance level across the 4 scanning sessions ($t(39) = 11.22$; $p<0.001$, $d= 1.76$). In addition, accuracy scores improved over time, with higher accuracy in the last than in the first sessions ($F(3, 37) = 14.76$; $p<0.001$, $\eta_p^2=0.28$). Additional evidence of successful acquisition was provided by a decrease in RTs over time, with higher RTs in the first than in the last sessions ($F(3, 37) = 27.08$; $p<0.001$, $\eta_p^2=0.41$). This was associated with increased PD for CS+ than CS- while participants were rating loss expectancy ($t(39) = 2.41$; $p=0.02$,

d= 0.16), and after they received the US (t (39) = 3.36; p=0.002, d= 0.22) across the 4 sessions. These results are shown in Figure 5.4.

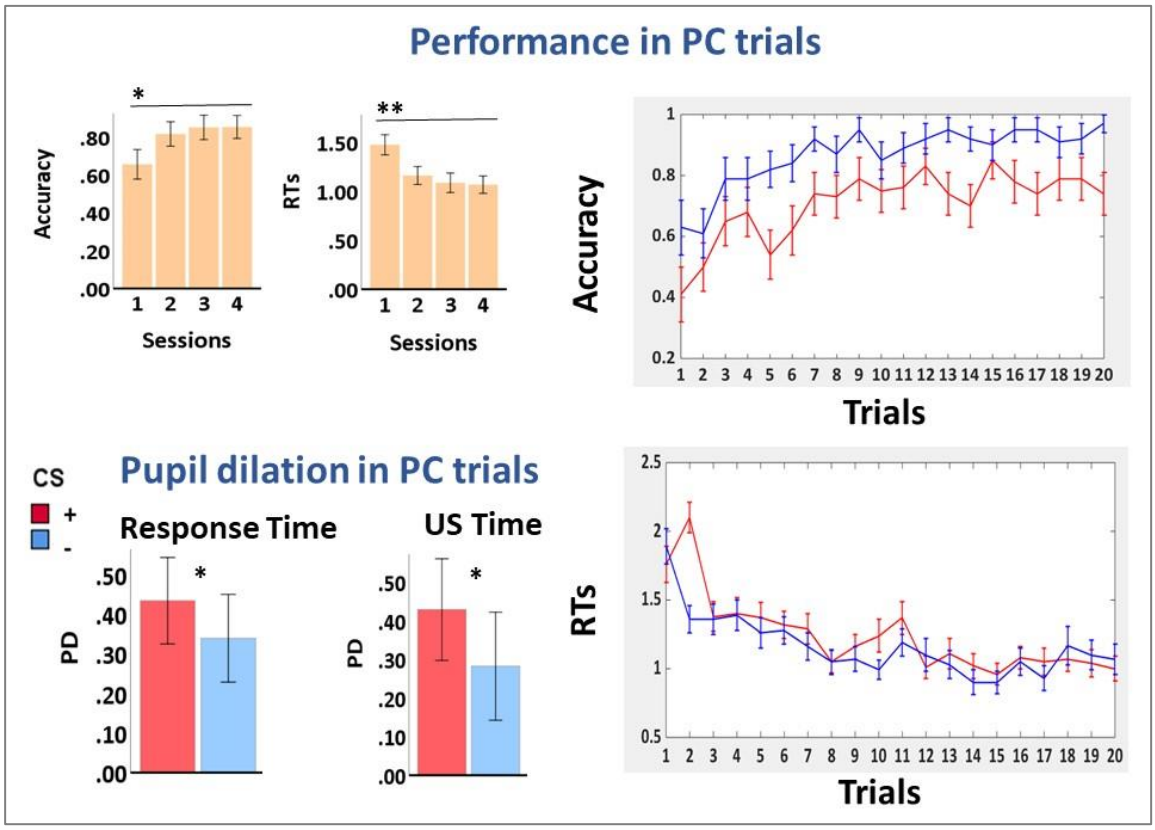


Figure 5.4. Experiment 2: Learning performance and pupil diameter (PD) during Pavlovian Conditioning (PC) trials, distributed across the 4 scanning sessions (10 trials within each session: 5 CS+ and 5 CS-). Accuracy was measured by dividing the number of correct answers for the number of stimuli within each condition, separately for CS+ and CS-, within each session. The RTs were also measured within each condition, separately for CS+ and CS-, within each session. Both accuracy and RTs were averaged across CS+ and CS- in the top left plot. PD measures at response time and when the US was delivered were z-scored transformed, by subtracting from the raw data the mean PD within each session across conditions, and dividing it by its standard deviation. Error bars represent ±2 SEM. *, p<0.05. Abbreviations: RTs, Reaction times; PD, pupil diameter.

Successful visual (VG) and auditory (AG) fear generalization to new exemplars was measured as in the PC trials. On average, participants generalised threat to new exemplars, evident in higher accuracy scores than chance level, in both VG (t (39) = 7.61; p<0.001, d= 1.20) and AG (t (39) = 8.14; p<0.001, d= 1.47). We also observed a decrease in RTs throughout the scanning sessions, with lower RTs in the second than in the first sessions of VG (t (39) = 12.64; p=0.001, d=0.16) and AG (t (39) = 30.68; p<0.001, d= 0.31). These results are shown in Figure 5.5. However, we did not find significant differences in PD between GS+ and GS-, either while participants were seeing images of/ listening the sounds of new exemplars or while they were providing their ratings (p>0.05).

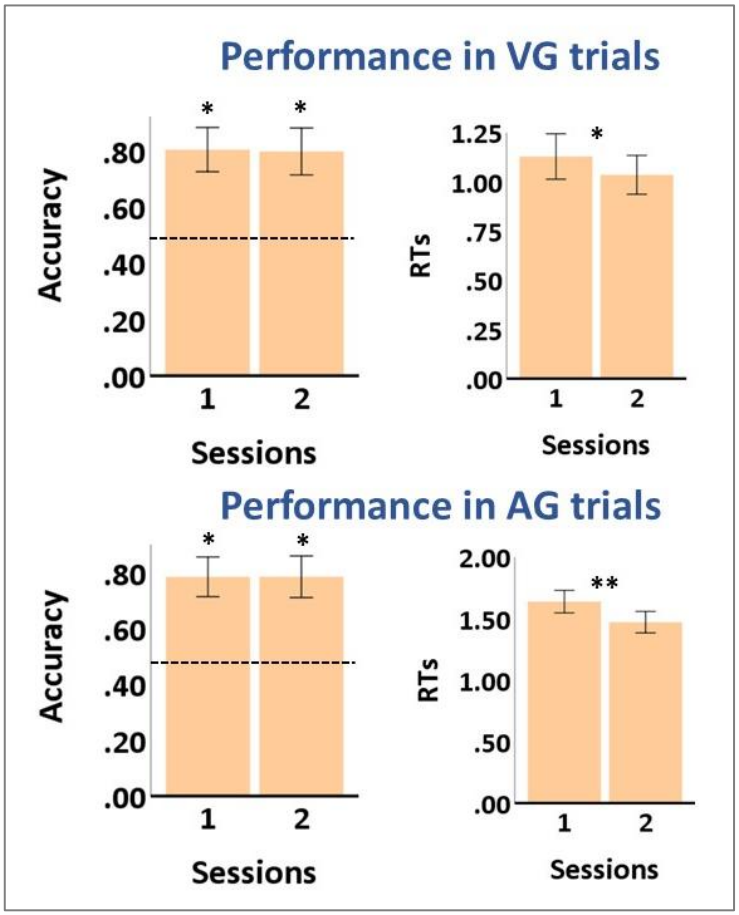


Figure 5.5. Experiment 2: Participants performance during Visual (VG) and Auditory (AG) generalisation trials. Accuracy was measured by dividing the number of correct answers for the number of stimuli within each condition, then averaged across GS+ and GS-, within each session. The RTs were measured in the same manner. The dash line represents the chance level. Error bars represent ±2 SEM. *, p<0.05. **, p<0.001. Abbreviations: GS, generalisation stimuli; VG, visual generalisation; AG, auditory generalisation; RTs, Reaction times.

Valence and arousal ratings between threatening than not threatening stimuli, and STAI_T

Further evidence that the conditioning and generalisation occurred were found in the valence and arousal rating task. Particularly, we observed differences in valence ratings between CS+ than CS- ($t(39) = -5.36$; $p<0.001$, $d= 1.73$) and between GS+ than GS- ($t(39) = -3.48$; $p=0.001$, $d= 1.12$), with lower valence scores in CS+/GS+ than CS-/GS-. Valence ratings were highly correlated across participants between CS+ and GS+ (Spearman's $\rho= 0.72$, $p<0.001$) and between CS- and GS- (Spearman's $\rho= 0.57$, $p<0.001$). Arousal ratings also differed between CS+ than CS- ($t(39) = 2.33$; $p=0.03$, $d= 0.51$) and between GS+ than GS- ($t(39) = 2.22$; $p=0.03$, $d= 0.50$), with higher arousal scores in CS+/GS+ than CS-/GS-. Arousal ratings were highly correlated between CS+ and GS+

(Spearman's ρ= 0.59, p<0.001) and between CS- and GS- (Spearman's ρ= 0.84, p<0.001). Finally, STAI_T was negatively correlated with valence ratings in CS+ (Spearman's ρ= -0.36, p=0.02) and in GS+ (Spearman's ρ= -0.31, p=0.04), and positively with arousal ratings in GS+ (Spearman's ρ= 0.41, p=0.009). These findings are shown in Figure 5.6.

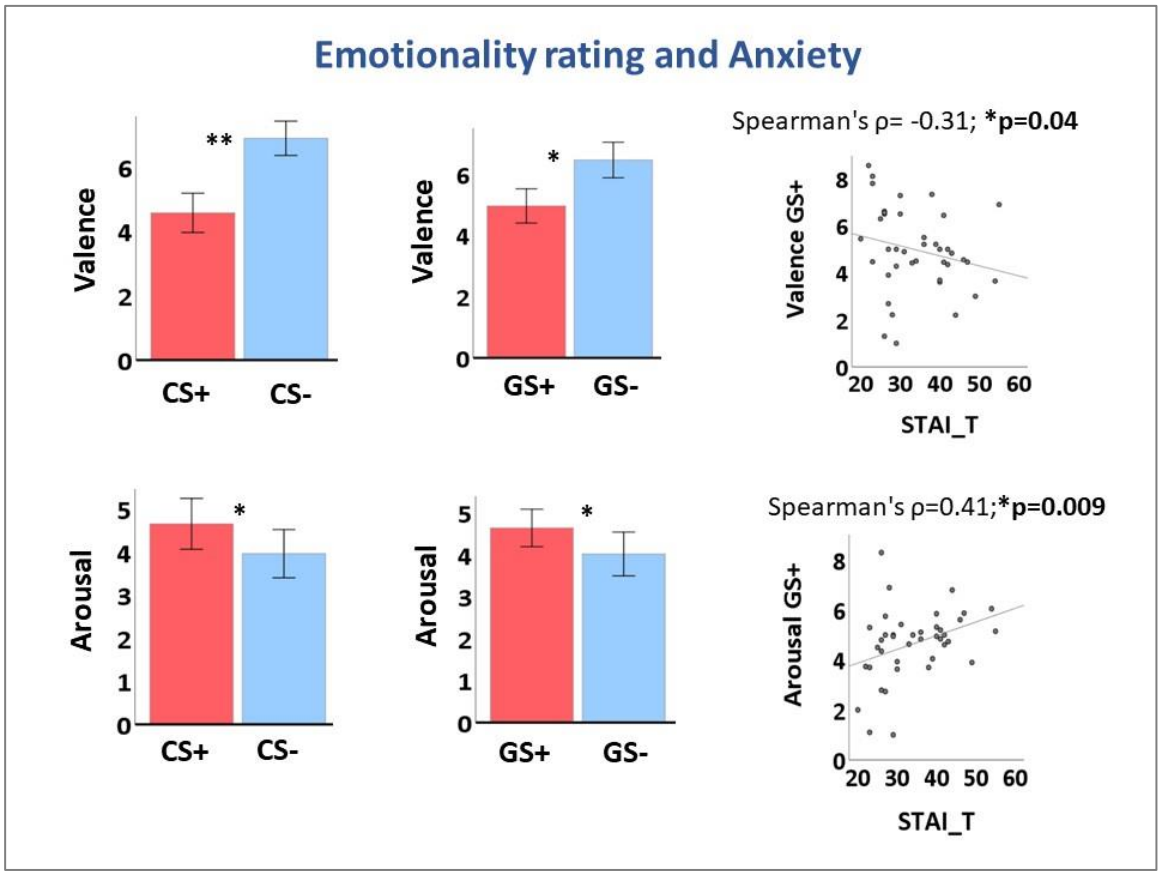


Figure 5.6. Experiment 2: differences in valence (top) and arousal (bottom) ratings between CS+ and CS-, and GS+ and GS-, averaged across participants. On the right, Spearman r between STAI_T and Valence GS+, and STAI_T and arousal GS+, averaged across participants. Error bars represent ±2 SEM. **, p< 0.001. *, p<0.05. Abbreviations: CS, Conditioned stimuli; GS, generalisation stimuli.

No differences in similarity perception among threatening stimuli

In Experiment 1, our manipulation check revealed lower dissimilarity within category (e.g., within mammals, within birds) than between categories (e.g., between mammals and birds) ($F(2, 18) = 56.10$, $p<0.001$, $\eta_p^2=0.75$). Crucially, there was no difference in similarity perception between threatening and not-threatening stimuli after conditioning ($F(1, 19) = 2.61$, $p=0.12$, $\eta_p^2=0.12$). This result is shown in Figure 5.7.

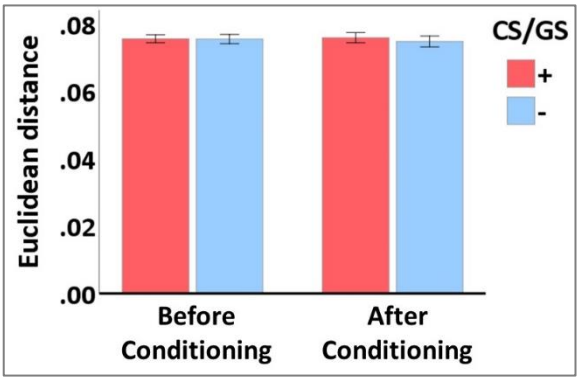


Figure 5.7. Experiment 1: Average dissimilarity between subordinate-level categories of threatening (CS+ and GS+) and not threatening (CS- and GS-) stimuli, before and after conditioning. Dissimilarity was measured as average Euclidean distance between subordinate-level categories of threatening (CS+ and GS+) and not threatening (CS+ and GS+) visual stimuli. Error bars represent ±2 SEM.

Neuroimaging data

Manipulation check: Univariate differences between GS+ and GS-

In VG, we observed significantly higher activation for GS+ than GS- in the right OFC, Insula, and MiFG ($p_{FWE}<0.05$), but not significant differences for the opposite contrast (i.e., GS+< GS-) in the ROIs mask. The activation in the right OFC during GS+ was significantly correlated with the valence GS+ ratings (Spearman's ρ=-0.37, p=0.019). In AG, we found that the right Thalamus and right MiFG were more activated during GS+ than GS-, and that the left Putamen was more activated for the opposite contrast (i.e., GS+< GS-) ($p_{FWE}<0.05$). These findings are shown in Table 5.1 and Figure 5.8.

Analysis	T Contrast	p FWE	K	p unc	T	x	y	z	Label
VG	CS+/ GS+ > CS-/ GS-	0.005	21	<0.001	4.01	24	47	-13	OFC R
		0.018	25	<0.001	3.51	33	-22	14	Insula R
		0.053	22	0.001	3.24	36	17	44	MiFG R
AG	CS+/ GS+ > CS-/ GS-	0.025	16	<0.001	3.52	6	-7	23	Thalamus R
		0.044	12	0.001	3.29	21	-7	38	MiFG R
	CS+/ GS+ < CS-/ GS-	0.048	26	0.001	3.26	-27	-4	-1	Putamen L

Table 5.1. Experiment 2: Differences in BOLD signal change between GS+ and GS- during visual and auditory generalisation in the ROIs mask. Results were Bonferroni corrected for multiple comparisons using pFWE < 0.05. Small volume correction was applied in the analyses. Abbreviations: OFC, Orbitofrontal cortex; MiFG, Middle frontal gyrus; R, right; L, left.

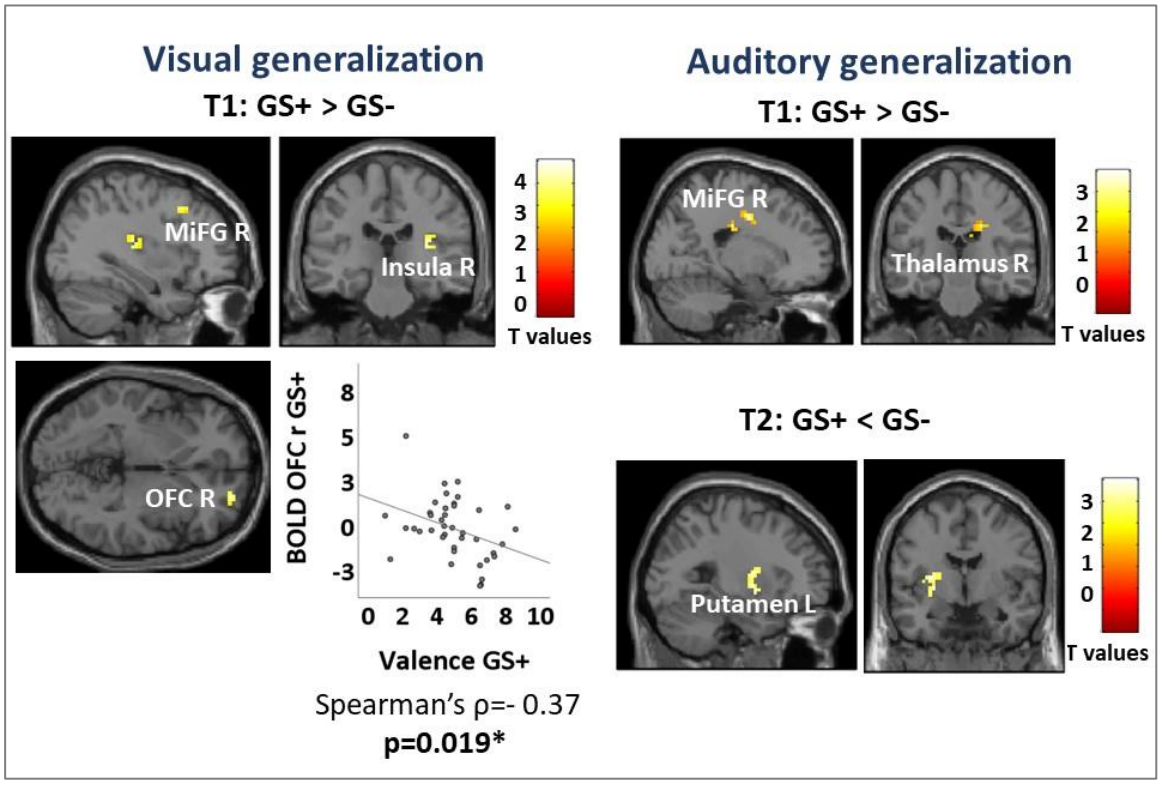


Figure 5.8. Experiment 2: differences in BOLD signal change between GS+ and GS- across sessions, during visual (left) and auditory (right) generalisation in the ROIs mask. Top: brain regions associated with higher activation for GS+ than GS- conditions. Bottom: brain regions associated with lower activation for GS+ than GS- conditions. Results were corrected for multiple comparisons using pFWE < 0.05. Small volume correction was applied in the analyses. We also computed the Spearman's correlation between brain activations and emotional ratings to GS+ and GS-. Abbreviations: Abbreviations: OFC, Orbitofrontal cortex; MiFG, Middle frontal gyrus; R, right; L, left.

RSA: neural generalisation across sensory modalities

In the whole sample, we did not observe differences in neural similarity in any ROI between GS+ and GS- in visual modality. Thus, we explored whether the same effect was present only in participants who generalised (n=31). As previously mentioned, successful generalisation was defined as an accuracy score higher than chance level (50%) across GSs computed separately for the VG and AG condition. We found higher neural similarity (lower dissimilarity) in GS+ than GS- in the bilateral FG (left: t (30) = -2.60, p=0.01, d= 0.05; right: t (30) = -2.15, p=0.04, d= 0.05), bilateral ITC (t (30) = -2.31, p=0.03, d= 0.04), and bilateral DMPFC (t (30) = -2.24, p=0.03, d= 0.05). These findings are shown in Table 5.2 and Figure 5.9.

ROIs	Conditions		Statistics		
	GS+	GS-	t	p	d
FG L	0.74 ± 0.12	0.76 ± 0.11	-2.60	0.01*	-0.05
FG R	0.71 ± 0.13	0.73 ± 0.12	-2.15	0.04*	-0.05
ITC	0.87 ± 0.07	0.89 ± 0.06	-2.31	0.03*	-0.04
DMPFC	0.90 ± 0.06	0.91 ± 0.05	-2.24	0.03*	-0.05

Table 5.2. Experiment 2: Effect of aversive conditioning on neural dissimilarity during visual generalisation. Difference in neural dissimilarity (measured as correlational distance) among conditions. The dissimilarity within GS+ and within GS- were calculated by averaging the dissimilarity within GS+ and within GS- across sessions, for each participant. These were entered as dependent variables in paired t tests, one for each brain cluster ($p<0.05$). *, $p<0.05$. Abbreviations: dissimilarity GS+, within GS+; GS-, within GS-; FG, Fusiform gyrus; ITC, Inferior temporal cortex; DMPFC, Dorsomedial prefrontal cortex; L, left; R, right.

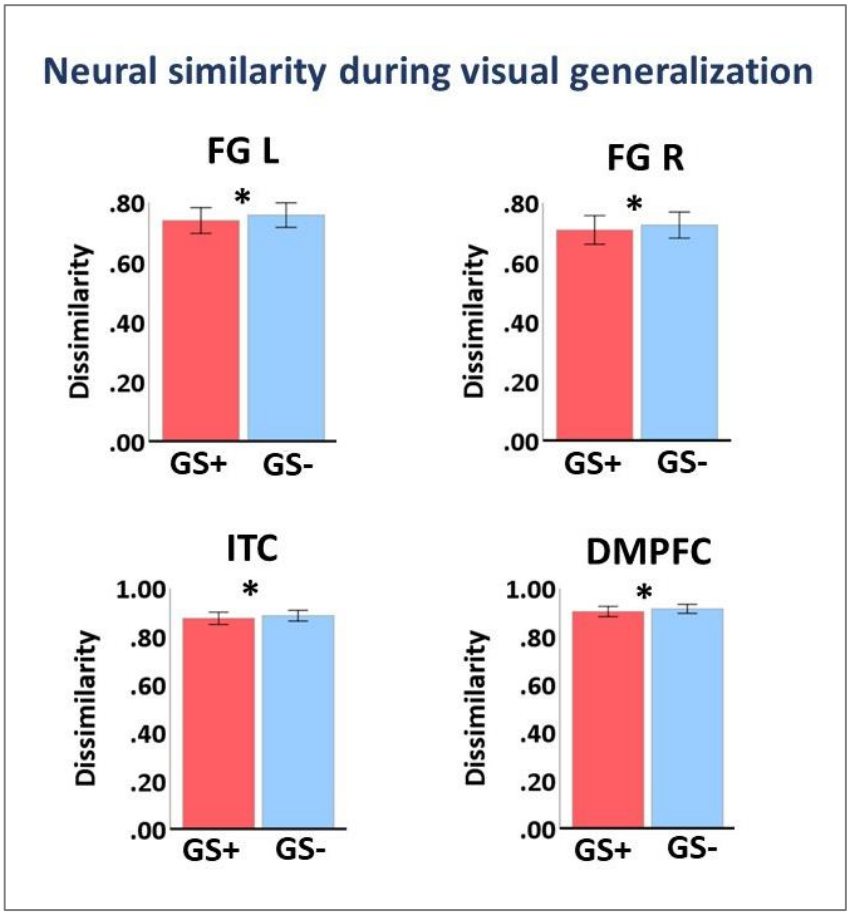


Figure 5.9. Experiment 2: Differences in correlational distance during visual generalisation within GS+ and within GS- in different brain regions, including the bilateral FG, ITC and DMPFC. The dissimilarity within GS+ and within GS- were calculated by averaging the dissimilarity within GS+ and within GS- across sessions, for each participant. These were entered as dependent variables in paired t tests, one for each brain cluster ($p<0.05$). *, $p<0.05$. Abbreviations: dissimilarity wGS+, within GS+; wGS-, within GS-; FG, Fusiform gyrus; ITC, Inferior temporal cortex; DMPFC, Dorsomedial prefrontal cortex; L, left; R, right.

However, we did not find any significant difference between auditory GS+ and GS-, either in the whole sample or in the subsample with only participants that generalised, pointing at equal similarity between auditory GS+ and auditory GS-. However, in the entire sample, we observed higher neural similarity across sensory modalities in GS+ than GS- in the bilateral Insula (t (39) = -2.37, p=0.02, d= -0.10), that is, higher similarity between visual and auditory GS+ than between visual and auditory GS-. This result is shown in Figure 5.10. The same effect was observed in the subsample of participants that generalised, although the effect was smaller (t (30) = -2.03, p=0.05, d=-0.09). However, our manipulation check of higher similarity between the image and the sound of a mammal (i.e., GS+) than between the image of a mammal and the sound of a bird (i.e., vGS+/aGS-), or between the image of a bird and the sound of a mammal (i.e., (vGS-/aGS+), revealed not significant results. We did not observe any of these effects in the other ROIs.

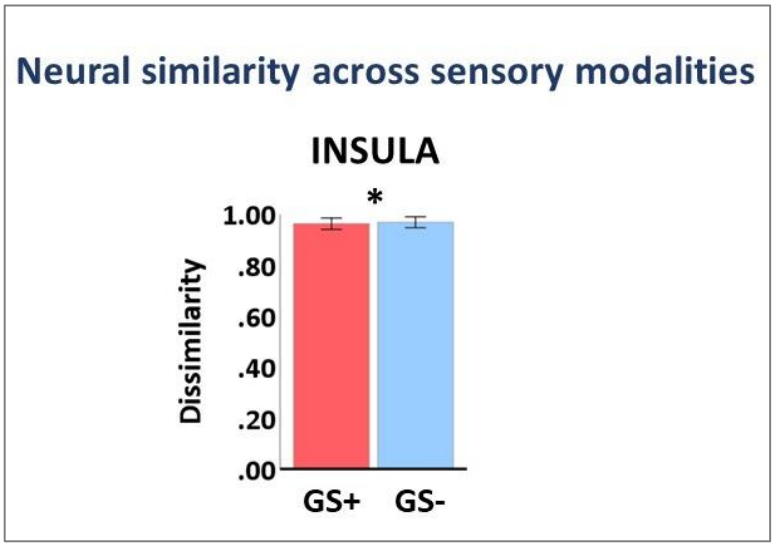


Figure 5.10. Experiment 2: Differences in dissimilarity (measured as correlational distance) between visual-auditory GS+ and visual-auditory GS- in the bilateral insula. The dissimilarity between visual-auditory GS+ and visual-auditory GS- calculated by averaging the dissimilarity within GS+ and within GS- across sessions, for each participant. These were entered as dependent variables in paired t tests, one for each brain cluster ($p<0.05$). *, $p<0.05$. Abbreviations: GS+, dissimilarity between visual and auditory GS+; GS-, dissimilarity between visual and auditory GS-.

5.4 Discussion

In a series of experiments, we investigated the effect of aversive conditioning on behavioural and neural measures of similarity between images and vocalisations of animals. We reported three findings. First, aversive conditioning increased the neural similarity between threatening visual stimuli, and between visual and auditory stimuli. The former effect was observed in the occipitotemporal cortex

and in the DMPFC in participants that generalised, the latter in the bilateral insula in the entire sample. Second, after conditioning, participants rated threatening stimuli as more negative and arousing than not threatening stimuli, and these ratings correlated with trait anxiety scores. Finally, despite the increase in neural similarity and changes in emotionality ratings after aversive conditioning, we did not observe any change in similarity perception among images of animals in an independent sample. Below we discuss the implications of these findings.

Increased neural similarity across sensory modalities after aversive conditioning

In experiment 2, we found that aversive conditioning is associated with a strengthening in neural similarity in GS+ than GS- in visual modality, and across sensory modalities. Part of our condition manipulation was successful, evident in differences in emotionality ratings and univariate activations after the conditioning between threatening and not threatening stimuli. Similarly to Levine et al. (2021) that found increased neural similarity in the superior frontal gyrus and right temporal pole between the CS+/GS+ superordinate- level categories (e.g., animals), we included images of exemplars from different subordinate-level categories to explore the same effect between CS+/GS+ subordinate-level categories (Levine, Kumpf, Rupprecht, & Schwarzbach, 2021). The novelty of our study was testing this effect across visual and auditory sensory modalities.

We observed increased neural similarity in visual modality in the occipitotemporal cortex, replicating previous finding (Dunsmoor et al., 2013), and in the DMPFC for the first time. The occipitotemporal cortex is involved in neural representation of visual stimuli in semantic categories, as previously shown (Charest et al., 2014; Kriegeskorte, Mur, Ruff, et al., 2008). The DMPFC has been involved in emotion regulation strategies by cognitive reappraisal during emotional learning (Buhle et al., 2014; Kohn et al., 2014). In addition, previous neuroimaging studies showed that brain activations to the CS+/ GS+ vs CS-/GS- in different brain regions, including DMPFC, decreased as the presented GS diverges from CS+ (for reviews see (Etkin, Egner, & Kalisch, 2011; Sehlmeyer et al., 2009). Lissek et al. (2014) proposed it as a neurobiological mechanisms underlying the behavioural overgeneralization seen in anxiety patients (Lissek et al., 2014), and it is consistent with the view of the DMPFC as neural hub that influences the activity in other brain

regions when threats are unpredictable (Wheelock et al., 2014). We speculate that the DMPFC encodes the threat value associated with the GS+ stimuli, and modulates the neural representations of threatening semantic category in the ventral-visual stream via thalamus projections. This might be the result of local (within a brain region) and distant (between brain regions) synchronisation processes, underlying higher similarity among threatening stimuli (Grandjean et al., 2008). More specifically, a stable mental representation (of the stimuli in GS+, for example) emerges from the patterning of neuronal assemblies in the ventral-visual stream, which encoded the 'relevance' of the stimulus (through thalamus-amygdala connections), and it is exchanged, through functional coupling via thalamus, with the DMPFC. This neural representation of the new 'threatening' category pattern is shared among threatening stimuli, and might explained their higher neural similarity. This mechanism is functional to survival, as it supports a fast 'fight or flight' action in front of new exemplars that more resembled the CS+ category (Levine et al., 2018).

In support to this, we also observed increased similarity between visual and auditory GS+ than GS- in the bilateral insula. Our findings are in line with those from Onat et al. (2015), showing a strengthening in neural similarity in the insula between CS+ and US in visual modality (Onat & Büchel, 2015). As part of the 'salience network', the insula forms a mental representation of the internal homeostatic state, activated in front of emotionally relevant stimuli via bottom-up connections of the insula with the rest of the body (Lindquist et al., 2012; Menon & Uddin, 2010). With this goal, it integrates aspects of bodily reactions triggered by external and internal stimuli, encoded in different neuronal subpopulations (e.g., cardiac, respiratory), through synchronisation process (Sander et al., 2018); it evaluates them and maps them into different bodily states representations (Onat & Büchel, 2015). The role of insula in fear generalisation across sensory modalities is further supported by clinical data showing that patients with anxiety disorders exhibit signs of impaired visceral and bodily processing (Paulus & Stein, 2006), and that specific phobias are associated with abnormal insular responses (Etkin & Wager, 2007).

In line with the findings of experiment 2, we observed that trait anxiety scores were negatively and positively correlated with GS+ ratings of valence and arousal, respectively, after conditioning. To our knowledge, no previous studies reported

an association between ratings of emotional dimensions of GS+ and trait anxiety in healthy controls. Wiemer et al. (2021) found similar results, but for emotionality ratings of GS- only (Wiemer, Rauner, Stegmann, & Pauli, 2021), suggesting that anxiety may be related to increased responses to safety stimuli (Duits et al., 2015). In their study, participants were asked either to up-regulate or down-regulate their emotions by reappraisal, while emotional and neutral faces were shown, some of which were paired with a scream as US and served as CS+. One explanation underlying this discrepancy might be because uncertainty might worsen fear generalisation in participants with high trait anxiety. Indeed, authors reported higher accuracy in memory performance for CS+ than CS-, suggesting higher uncertainty among participants in the latter than former condition. Uncertainty might increase anxiety, especially in participants with high trait anxiety scores, to the point that safe stimuli are misjudged as threatening. Conversely, in our study, given the high number of GS trials, uncertainty was almost absent on average, as evident in the high accuracy in the first session of VG and AG. Participants with high trait anxiety scores perceived the threatening stimuli as more negative and arousing, suggesting that anxiety may amplify negative emotions, resulting in difficulties in emotion regulations previously reported in anxious individuals (Cho, White, Yang, & Soto, 2019). Specifically, anxious individuals struggle in early stage of emotion processing (Liu, Wang, & Li, 2018), leading to choosing ineffective emotional regulations strategies, and even when they are effective, a more intense internal emotional response may nullify the benefits of that strategy (Soto et al., 2012). This is also in line with a previous finding that undergraduates with high trait social anxiety perceive negative mental images in feared situations as less controllable than less anxious students (Moscovitch et al., 2013).

However, the data on neural similarity across sensory modalities present some limitations, as we did not observe an increase in neural similarity in auditory GS+, and our manipulation check for the RSA across sensory modalities did not reveale significant results. In other words, the neural similarity of the same condition across different sensory modalities (e.g., visual and auditory GS+) was not significantly higher than the neural similarity between different conditions across different sensory modalities (e.g., vGS+/aGS-). One of the reasons might be that Aversive conditioning acts by increasing the similarity within the conditioned modality (i.e., visual), and across conditioned and not-conditioned modalities,

rather than also within the not conditioned (i.e., auditory) one. An alternative explanation might be related to the low emotional impact of the US. This is also supported by emotionality ratings, showing that, despite the differences in valence and arousal ratings between GS+ and GS-, both conditions present neutral valence and arousal, as both measures ranged between 4 and 6. In addition, we did not observe any difference in pupil diameter between GS+ and GS- during visual and auditory generalisation, but only during Pavlovian conditioning trials, suggesting a decreased emotional impact over time. Given that PC trials were fewer compared to GS trials, and half of them were presented at the beginning of first session (both in auditory and visual generalisation blocks), participants were naïve while seeing half of the PC trials, resulting in higher emotional impact than in GS trials. Also, the slow-event related design (long inter-stimulus intervals), and possibly secondary appraisals (i.e., implications) of the GS+ stimuli might have contributed to the reduced emotional effect, as participants were aware they would have been reimbursed for their time. This is also in line with neuroimaging data, which did not reveal any difference in activation between GS+ and GS-, for example, in the Amygdala. Delgado et al. (2008) found that thinking about something calming in nature, while viewing a conditioned stimulus (CS+) predicting an aversive electric shock, reduces the triggered conditioned responses (i.e. skin conductance and amygdala activity) (Delgado, Nearing, LeDoux, & Phelps, 2008). The same might be true in our experiment, when participants reappraised their fear of losing money with a more reassuring thought (i.e., reimbursement).

Taken together, these findings suggest that aversive conditioning occurred and fear generalised between CS+/GS+ subordinate-level categories, as supported by behavioural (accuracy, RTs, valence and arousal ratings) and neurophysiological (PD, BOLD signal) data. We also provided some light on the neural mechanisms of fear generalisation, showing that it is associated with an increase in neural similarity between stimuli in GS+ category within the conditioning modality, as well as across conditioned and not- conditioned sensory modalities (i.e., visual and auditory). Despite these changes, and the effect of trait anxiety scores on emotionality ratings of GSs, the perception of similarity among them did not change following conditioning.

Similarity perception does not change after aversive conditioning

In experiment 1, we did not find increased similarity between subordinate-level categories in the threatening than not-threatening conditions following aversive conditioning, despite the differences between them in neural similarity observed in experiment 2. This might suggest that aversive conditioning does not change similarity perception, but it affects the neural representation of threatening stimuli, which are semantically similar to the CS+. From an evolutionary perspective, a strengthening in the neural similarity among stimuli that predict the same outcome (i.e., possible threat) is more efficient, as they can be grouped together in the same (emotional) category, encoded as regularities in neural patterns (Onat & Büchel, 2015), rather than as separate events. However, in behaviour, when participants are asked to explicitly judge inter-stimulus similarity (rather than rating money loss expectancy of each stimulus), semantic features might be more relevant. Participants arranged the stimuli in the arena according to their semantic similarity, across the different levels of categories, before and after the conditioning. These results demonstrate that aversive conditioning might not affect similarity perception, but has an influence on emotional processing and categorisation. This fact is supported by differences in emotionality ratings between threatening and not-threatening stimuli after conditioning, their correlation with trait anxiety scores, and by experience-dependent changes in PD and BOLD signal among threatening stimuli in experiment 2.

An alternative explanation might be that secondary negative reinforcements as US in an experimental setting does not have a high emotional impact on participants. Participants mentally represent threatening stimuli as more negative than not-threatening, but they do not experience conscious emotional threat. However, previous studies using similar US (Laufer & Paz, 2012; Lissek et al., 2009) detected behavioural changes in fear generalisation, with the difference that they used simple visual or auditory GSs (i.e., shapes, tones), which differed along perceptual dimensions (e.g., size, pitch). In our study, we included an additional (not perceptual) feature into the similarity space, which might be as relevant as visual and emotional dimensions in judging inter-stimulus relationships (Martina Riberto et al., 2021).

Limitations

Our study presents several limitations that can be addressed in future work. First, our materials were limited. We studied only aversive, and not reward conditioning. We chose this because only the former is relevant for overgeneralisation bias in anxiety disorder. It would be interesting in future studies to examine whether reward conditioning is also associated with an increase in neural similarity across sensory modalities. Second, the US may not have had a high emotional impact, as shown in the neurophysiological and neuroimaging data. This might also explain the small effect sizes of differences in neural similarity between GS+ and GS- detected in experiment 2, within and across sensory modalities. Future studies should explore the same research question by adopting a primary rather than secondary negative reinforcements (Delgado, Labouliere, & Phelps, 2006). Finally, we cannot infer any causal role of aversive conditioning on neural similarity from our study. Future studies could use TMS to further explore this aspect of the findings.

Conclusions

In conclusion, we observed a strengthening in the neural similarity structure of the stimuli that resemble the CS+ (i.e., GS+) after aversive conditioning. This effect was detected between visual GS stimuli in the occipital-temporal cortex and in the DMPFC, and between visual and auditory GS+ in the insula. These data were paralleled by experience-dependent changes in PD, BOLD signal and emotionality ratings. Despite the differences in neural similarity, the similarity perception between threatening stimuli remained unchanged. We speculate that increased neural similarity might be the result of local and distance synchronisation processes between brain regions involved in fear generalisation. This might be an efficient and functional mechanism, which may become maladaptive in individuals with anxiety disorders.

6. Chapter: General discussion

In this thesis, our aims were twofold. First, we were interested in when and why two emotional experiences are similar to each other, and second, whether the neural representations of emotional stimuli were more similar to each other than those of neutral stimuli. We assumed a similarity space that comprises of several integrated visual (e.g., luminance, contrast), semantic (e.g., taxonomy, theme) and emotional (e.g., valence, arousal, appraisals) dimensions. In this space, stimuli are displaced according to the similarity between them along these dimensions, with short distances reflecting high similarity. We expected emotional features to be the most influential in perceived similarity, because it enables individuals to differentiate emotionally relevant stimuli from those that are not as function of survival, following primary appraisals. According to this, we predicted emotional stimuli (e.g., fearful and disgust faces; threatening animals; car accidents and poverty scenes) to be placed in the similarity space closer to each other than neutral stimuli (e.g., neutral faces; not threatening animals; scenes of laundry and phone call), as they shared low scores in valence and high scores in arousal, and thus more salient. We also expected this to be underpinned by higher similarity in the neural activation patterns of emotional than neutral stimuli. We tested our hypotheses using two similarity judgements tasks (i.e., pairwise ratings and MA tasks) and several databases of stimuli that differed in visual complexity (e.g., pictures of animals, complex scenes) and along emotional dimensions (e.g., fearful faces; scenes of car accidents). Specifically, we selected both stimuli that were emotional *per se* in **chapter 3** (e.g., fearful and disgusted faces) and **chapter 4** (e.g., pictures of car accidents, poverty scenes), and in **chapter 5** neutral images and sounds of animals stimuli that assumed an emotional value following fear conditioning.

Specifically, in **chapter 3** we found that, even though two expressions of the same person were physically more similar to each other than the faces of two different individuals who expressed the same emotion, participants judged both types of face pairs to be just as similar to each other. This suggests an important role of emotion on similarity perception that goes beyond the objective visual reality of the face. While emotion mattered to similarity perception, this was not because participants perceived stimuli high in emotion to be more similar than those low in

emotion. In fact, emotional faces were perceived to be just as similar as neutral faces. It was the presence of emotion that mattered: when emotionality mismatched (i.e., EE_DF and EN), faces were perceived as least similar, and when it matched (i.e., EE and NN), as most similar. We replicated the same findings in the behavioural experiments in **chapter 4**, wherein emotional mismatch mattered a great deal. Specifically, EN pairs were consistently judged as least similar compared to conditions of emotional match (e.g., EE and NN) and to those wherein negative emotions were qualitatively mismatched according to a 'basic emotions' approach, but similar in valence and arousal, assuming a dimensional perspective (e.g., E12, N12). Also, in **chapter 5** (*experiment 1*) pairs in EN were judged as the least similar.

Nevertheless, emotional dimensions were the most influential in the similarity space, only when thematic similarity was not controlled (**chapter 4**, *experiment 1*). This resulted in higher similarity ratings for emotional than randomly-selected neutral pictures. When semantic similarity was controlled (**chapter 4**, *experiments 2-3*) participants subjectively perceived negative emotional stimuli to be just as similar as neutral stimuli. Ratings clustered according to the four categories, suggesting that the semantic meaning of each picture – not negative emotion - was the most relevant feature. The same pattern of results has been found in **chapter 5** (*experiment 1*), wherein aversive conditioning was not associated with increase in similarity between threatening and not threatening visual stimuli.

Despite this equivalence in similarity perception, emotional dimensions were more influential in the neural similarity space, resulting in higher similarity among the neural representations of emotional compared to neutral stimuli. In particular, in **chapter 4** (*experiment 3*) we observed a strengthening in neural similarity in brain clusters located in the ventral visual processing stream underlying semantic processing and categorisation, and in regions involved in affect representations (i.e., precuneus and anterior insula) and modulation (i.e., dorsal anterior cingulate cortex). Consistently, in **chapter 5** (*experiment 2*) we found increased neural similarity in the occipitotemporal cortex among threatening visual stimuli, and in the insula between visual and auditory stimuli that predicted a negative outcome (i.e., money loss), following aversive conditioning. We discuss the implications of these results below.

Increased neural similarity between emotional than neutral stimuli.

In **chapters 4** (*experiment 3*) and **5** (*experiment 2*) we observed that the neural representations of emotional stimuli were more similar than those associated with neutral stimuli. We observed this effect using visually complex stimuli, which evoked negative emotions per se', and simple stimuli, which were emotional in an experience-dependent fashion. In both experiments, negative emotions were associated with a strengthening in neural similarity. This might suggest that the brain encodes more similarly stimuli that share an emotional value compared to those rated as non-emotional. In other words, neural similarity as measured by RSA might inform us about how different neural populations respond to the experimental conditions, and in line with this, are clustered in the (neural) similarity space. High proximity between neural representations suggest that the stimuli are grouped together as they share a neural pattern which emerges as stable representation from underlying dynamic system (e.g., signalling between neurons or between brain regions). The neural space is defined according to the features that are functional to discriminate the stimuli (e.g., low-level visual features, semantic meaning, valence) in a parsimonious way. Each feature is encoded in brain regions involved in computing different aspects of the stimuli (e.g. the ventral-visual stream carries out the semantic processing; the precuneus represents the valence). In effect, we found increased neural similarity for emotional than neutral stimuli in occipitotemporal cortex (e.g., ITC, FG), which underpins semantic processing and categorisation (Clarke & Tyler, 2014; Iordan et al., 2015), and in brain regions involved in affect representations (i.e., precuneus, insula) (Giordano et al., 2021; Kim et al., 2017) and modulation (i.e., dACC and DMPFC) (Kohn et al., 2014; Saarimäki et al., 2018).

Emotion-dependent increases in neural similarity suggests that emotion might be the most influential dimension in the neural similarity space. According to this, we define what is bad or good for us, such that any representation of the world is an integrated product between emotion, perception and thought (e.g., 'That is a good thing') rather than isolated psychological events (e.g., 'That is a thing. I feel good') (Todd et al., 2020). Thus, 'bad things' might have similar neural representations in the precuneus because they share (negative) valence, as well as in regions involved in emotions regulation, as they might induce an emotional response that

is more difficult to suppress than neutral stimuli. In addition, 'bad things' might have similar neural profiles in high-level and low-level visual regions as a result of top-down neural modulation. Thus, irrespective of instructions or task demand, emotional similarity might be considered as a functional organising principle for the brain that helps to define which stimuli are more salient in the early stage of neural processing.

In **chapter 4** (*experiment 3*), we found that the similarity among the neural representations of emotional compared to neutral categories was higher, both in regions that also expressed similarity between neutral stimuli, and in unique regions. Specifically, the early visual cortex (EVC) and brain clusters responsible for affect representation and emotion regulation (i.e., precuneus, anterior insula and dACC) uniquely represented the similarity between emotional stimuli. Conversely, high-level visual regions involved in semantic processing and scenes categorisation (i.e., ITC, FFA, OPA and PPA) encoded the similarity between the two emotional and the two neutral categories, but more strongly in the emotional condition. OPA and PPA relate low-level visual features encoded in the EVC with the high-level aspects of the scene (R. A. Epstein & Baker, 2019), and like other ITC regions, may be modulated by regions that are sensitive to salience, such as anterior insula and dACC (Lindquist et al., 2012), resulting in higher similarity. Likewise, we may have observed stronger neural similarity between emotional categories in occipitotemporal regions (Kim et al., 2017) because of the influence of the precuneus, which is involved in valence representation and structurally connected with the ITC (Y.-H. Lin et al., 2020).

In **chapter 5** (*experiment 2*) we replicated previous studies in fear conditioning literature with visual stimuli (Dunsmoor et al., 2013; Levine et al., 2021; Onat & Büchel, 2015; R. M. Visser et al., 2011) and extended them across sensory modalities. Particularly, we observed a strengthening in neural similarity after fear conditioning among visual GS, which were semantically related to the CS+, in the ventral visual stream (e.g., ITC, FG) and in brain regions that carried out emotion-relevant computations, including the DMPFC. Furthermore, we found higher neural similarity between visual and auditory GS+ in the insula, which encoded a mental representation of the internal homeostatic state. This converging evidence might suggest that it is evolutionarily more important to integrate the emotional information in neural representations in order to increase the relevance of the

stimulus, rather than separately encode the meaning of the stimulus and its emotional value. This mechanism is adaptive, because it enables individuals to differentiate emotionally salient stimuli from those that are not, and might support broad generalisation between stimuli, which predict the same fitness-relevant outcome.

This increase in neural similarity might be the result of local (within a brain region) and distant (between brain regions) synchronisation processes (Grandjean et al., 2008) triggered by emotional stimuli. As we suggested in previous chapters, brain regions that carry out emotion-relevant computations (e.g., DMPFC, insula) encodes the emotional value associated with each stimulus and modulates their neural representations in the ventral-visual stream via feedback projections. A stable mental representation (e.g., GS+, scenes of car accident) emerges from the patterning of neuronal assemblies in the ventral-visual stream (e.g., ITC, PPA), which encoded the 'relevance' of the stimulus. This neural representation pattern is shared among emotional stimuli (from the same or emotionally similar category), and might explain their higher neural similarity. This mechanism is functional to survival, as it supports a fast 'fight or flight' action in front of new exemplars that more resembled previously encountered salient stimuli (Levine et al., 2018). In future studies, it would be worth testing differences in relatedness between brain data (from different regions) and computational models that simulate different aspects of the stimuli (e.g., visual features, tasks instructions, avoidance/approach), in order to provide further support to the emotional similarity hypothesis.

These neural findings have implications for research about the neurobiological correlates of semantic categorisation and generalisation. Previous model in semantic cognition literature (e.g., 'hub and spoke' model (Lambon Ralph, 2014)) focused on neutral stimuli, without considering the neurobiological mechanisms underlying emotional categorisation. We proposed that 'emotional spokes' (e.g., the precuneus) encoded the emotional dimensions (e.g., valence) of internal or external stimuli, which are important for forming a coherent mental representation of the world, as well as for planning appropriate emotion regulation strategies (Buhle et al., 2014; Kohn et al., 2014). According to this perspective, emotional similarity might represent a further dimension in a complex similarity space rather than a facet of semantic similarity. However, future research is needed to test

whether emotional spokes other than the hub are recruited during similarity judgements among emotional than neutral stimuli. From a clinical perspective, it would be worth exploring neural differences in similarity judgements between patients with semantic dementia and healthy controls. It is also important to understand the effect of emotion on neural similarity, because aberrant similarity perception could impact psychological well-being (Puccetti et al., 2021) and be relevant for the overgeneralisation bias in anxiety and posttraumatic stress disorders (Dunsmoor & Paz, 2015). After a traumatic event, patients may consider later experiences as similar to the original fearful one not because of their ostensible meaning, but due to their emotional similarity (Laufer et al., 2016). This might be the result of increased neural similarity among stimuli that resemble the fearful one after the traumatic experience. This adaptive mechanism may become maladaptive when old threat memories are inappropriately reactivated in secure situations, for example in individuals with high trait anxiety.

No differences in similarity judgements between emotional and neutral stimuli.

Despite the stronger neural similarity between emotional stimuli, participants judged them as equally similar to neutral stimuli, in contrast to our hypothesis. We expected that emotional feature was the most influential not only in the neural similarity space, but also during explicit similarity judgements. Thus, the increased neural similarity between emotional than neutral stimuli would be paralleled by higher similarity ratings for emotional stimuli. To test this in behavioral tasks, in **chapters 3, 4** and **5** (*experiment 1*) we used both visually simple stimuli (i.e., chapter 3: faces; chapter 5: images of animals) and more complex scenes (chapter 4). Using simple stimuli from the same semantic category allowed us to exclude semantic similarity as confounding factor. Conversely, complex scenes depicting events in a realistic context evoked higher emotional impact than simple stimuli, as shown by valence and arousal ratings in **chapter 4** (*experiments 1-2*) and **5** (*experiment 2*). However, when using more complex stimuli we had to control for several confounders (e.g., thematic similarity). Matching emotional and neutral complex scenes on semantic similarity is of paramount importance in similarity judgements tasks, as negative stimuli are often more thematically related than neutral stimuli. In **chapter 4** (*experiment 1*), these differences in thematic similarity

explained the higher similarity ratings among emotional than neutral pictures. Once semantic similarity was controlled (*experiments 2-3*), emotional and neutral pictures were rated as equally similar, and were arranged in a bidimensional similarity space according to the overall meaning of the pictures. These results suggested that participants considered the semantic dimension as the most relevant in the similarity space compared, for example, to the emotional features. This is consistent with the findings in **chapter 5** (*experiment 1*), wherein participants arranged threatening and not threatening stimuli according to their semantic rather than emotional similarity following aversive conditioning. This further proved that, when semantic similarity is controlled, the meaning of each picture was the most relevant feature, and thus, negative stimuli were not rated as more similar, rather equally similar to neutral stimuli. Aversive conditioning might have an influence on emotional processing and categorisation. This is revealed by differences in emotionality ratings between threatening and not threatening stimuli after conditioning, their correlation with trait anxiety scores, and by experience-dependent changes in PD and BOLD signal among threatening stimuli in **chapter 5** (*experiment 2*).

In **chapter 3**, emotional features were found to be as relevant as visual identity during explicit similarity judgements tasks. According to this, paired faces with different identity that expressed the same or similar emotions (i.e., Mr Hyde and Jack the Ripper) were rated as similar as faces with the same identity, but different emotional expressions (i.e., Mr Hyde vs Dr Jekyll). In other words, participants perceived Mr Hyde to be just as similar to Dr Jekyll (identity) as to Jack the Ripper (emotion). In the same experiment, we also observed that emotional mismatch (i.e., EN, EE_DF pairs) decreased perceived similarity. This suggests an important role of emotion on similarity perception that goes beyond the objective visual reality of the face. As proposed in previous studies (Jamin Brett Halberstadt & Niedenthal, 1997; Wegrzyn et al., 2017), the role of emotions in similarity judgements is evolutionary advantageous: poor discrimination among emotional expressions that have the same meaning (expressions of fear, for example) possibly would not endanger the individual.

However, when stimuli convey qualitatively different emotions, as in EE_DF (e.g., fearful and disgusted faces) and EN (e.g., emotional and neutral faces), small

dissimilarities can create large differences in similarity perception and action planning (e.g., fight or flight).

When and why two emotional experiences are similar to each other

To conclude, the pattern of findings across the experiments accords with claims that participants' similarity space comprises of several dimensions, including perceptual, semantic and emotional features. The relevance of each feature on the overall perceived similarity depends on the stimuli included in the experimental dataset and the control of possible confounding factors. Specifically, participants judged emotional stimuli as similar to each other, when they evoked the same basic emotion (e.g., fear). In **chapter 3** and **5** (*experiment 1*), emotionally matched conditions (e.g., chapter 3: EE and NN; chapter 5: CS+/GS+ and CS-GS-) were judged as the most similar compared, for example, to emotionally or visually similar conditions. In addition, two negative emotional experiences are judged as similar to each other when they are thematically-related (**chapter 4**: *experiment 1*). This is because negatively-valenced and highly arousing stimuli co-occur very frequently in the same theme or scenario, as the range of themes within negative and arousing stimuli (e.g. death, violence, car accidents, hospital scenes, and assaults) is reduced compared to those within neutral images. The fact of sharing emotional dimensions, such as valence and arousal, as in the context of the circumplex model of emotions, also make two emotional experiences similar to each other. In particular, in **chapter 4** (*experiments 2-3*) stimuli from the two emotional categories, which shared negative valence and high arousal, were perceived as more similar than stimuli from emotional and neutral categories (e.g., EN pairs). The same was valid for stimuli from the two neutral categories. The relevance of emotional features in the similarity space also depends on the experimental context (i.e., the set of stimuli included in the experimental situation). For example, when thematic similarity is not controlled, the emotional dimension might be the most relevant to group stimuli together. As a result, emotional stimuli were judged as more similar than randomly-selected neutral stimuli. When thematic similarity is controlled, participants relied more on semantic features. Conversely, the neural data in **chapter 4** (*experiment 3*) and **5** (*experiment 2*) also confirmed the relevance of emotional dimensions on neural similarity. This might be the result of local (within a brain region) and distant (between brain regions)

synchronisation processes, which are triggered by emotional experiences. A stable mental representation, which encoded the ‘relevance’ of the stimulus, emerges from the patterning of neuronal assemblies distributed in the brain. This neural representation pattern might be shared among emotional stimuli (from the same or emotionally similar category), and might explain their higher neural similarity.

References

Ahrens, L. M., Pauli, P., Reif, A., Mühlberger, A., Langs, G., Aalderink, T., & Wieser, M. J. (2016). Fear conditioning and stimulus generalization in patients with social anxiety disorder. *Journal of Anxiety Disorders, 44*, 36-46.

Alamia, A., VanRullen, R., Pasqualotto, E., Mouraux, A., & Zenon, A. (2019). Pupil-linked arousal responds to unconscious surprisal. *Journal of Neuroscience, 39*(27), 5369-5376.

Anglin, J. M. (1977). *Word, object, and conceptual development*: WW Norton New York.

Aviezer, H., Hassin, R. R., Ryan, J., Grady, C., Susskind, J., Anderson, A., . . . Bentin, S. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychological science, 19*(7), 724-732.

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology, 87*(2), 266.

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience, 12*(1), 1-23.

Barrett, L. F., & Bar, M. (2009). See it with feeling: affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1325-1334.

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition & Emotion, 15*(6), 713-724.

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science, 8*(1), 10-14.

Barsalou, L. W. (2017). What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia, 105*, 18-38.

Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., & Shinkareva, S. V. (2012). Decoding the neural representation of affective states. *NeuroImage, 59*(1), 718-727.

Bex, P. J., & Makous, W. (2002). Spatial frequency, phase, and the contrast of natural images. *JOSA A, 19*(6), 1096-1106.

Biondi, G., Franzoni, V., Li, Y., & Milani, A. (2016). *Web-based similarity for emotion recognition in web objects.* Paper presented at the Proceedings of the 9th International Conference on Utility and Cloud Computing.

Bozeat, S., Ralph, M. A. L., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia, 38*(9), 1207-1215.

Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition, 18*(2), 379.

Brooks, J. A., Chikazoe, J., Sadato, N., & Freeman, J. B. (2019). The neural representation of facial-emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences, 116*(32), 15861-15870.

Bruffaerts, R., Dupont, P., Peeters, R., De Deyne, S., Storms, G., & Vandenberghe, R. (2013). Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *Journal of Neuroscience, 33*(47), 18597-18607.

Bruner, J. S., & Austin, G. A. (1986). *A study of thinking*: Transaction publishers.

Buhle, J. T., Silvers, J. A., Wager, T. D., Lopez, R., Onyemekwu, C., Kober, H., . . . Ochsner, K. N. (2014). Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. *Cerebral cortex, 24*(11), 2981-2990.

Calder, A. J. (2003). Disgust discussed. *Annals of Neurology*.

Calhoun, V. D., Stevens, M. C., Pearlson, G. D., & Kiehl, K. A. (2004). fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms. *Neuroimage, 22*(1), 252-257.

Calvo, M. G., & Nummenmaa, L. (2008). Detection of emotional faces: salient physical features guide effective visual search. *Journal of Experimental Psychology: General, 137*(3), 471.

Caramazza, A., Hillis, A. E., Rapp, B. C., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cognitive neuropsychology, 7*(3), 161-189.

Carrasco, M., & Ridout, J. B. (1993). Olfactory perception and olfactory imagery: a multidimensional analysis. *Journal of Experimental Psychology: Human Perception and Performance, 19*(2), 287.

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS biology, 13*(6), e1002180.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences, 111*(40), 14565-14570.

Chavez, R. S., & Heatherton, T. F. (2015). Representational similarity of social and valence information in the medial pFC. *Journal of Cognitive Neuroscience, 27*(1), 73-82.

Chen, Q., Li, P., Xi, L., Li, F., Lei, Y., & Li, H. (2013). How do taxonomic versus thematic relations impact similarity and difference judgments? An ERP study. *International Journal of Psychophysiology, 90*(2), 135-142.

Chikazoe, J., Lee, D. H., Kriegeskorte, N., & Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature neuroscience, 17*(8), 1114.

Cho, S., White, K. H., Yang, Y., & Soto, J. A. (2019). The role of trait anxiety in the selection of emotion regulation strategies and subsequent effectiveness. *Personality and Individual Differences, 147*, 326-331.

Clarke, A., & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience, 34*(14), 4766-4775.

Cohen, A. O., Matese, N. G., Filimontseva, A., Shen, X., Shi, T. C., Livne, E., & Hartley, C. A. (2019). Aversive learning strengthens episodic memory in both adolescents and adults. *Learning & Memory, 26*(7), 272-279.

Collins, J. A., & Olson, I. R. (2014). Beyond the FFA: the role of the ventral anterior temporal lobes in face processing. *Neuropsychologia, 61*, 65-79.

Coutanche, M. N., Solomon, S. H., & Thompson-Schill, S. L. (2016). A meta-analysis of fMRI decoding: Quantifying influences on human visual population codes. *Neuropsychologia, 82*, 134-141.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences, 114*(38), E7900-E7909.

Damasio, A. (2003). Feelings of emotion and the self. *Annals of the New York Academy of Sciences, 1001*(1), 253-261.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural computation, 1*(1), 123-132.

Davis, R. G. (1977). Acquisition and retention of verbal associations to olfactory and abstract visual stimuli of varying similarity. *Journal of Experimental Psychology: Human Learning and Memory, 3*(1), 37.

Delgado, M. R., Labouliere, C. D., & Phelps, E. A. (2006). Fear of losing money? Aversive conditioning with secondary reinforcers. *Social cognitive and affective neuroscience, 1*(3), 250-259.

Delgado, M. R., Nearing, K. I., LeDoux, J. E., & Phelps, E. A. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. *Neuron, 59*(5), 829-838.

den Stock, J. V., Vandenbulcke, M., Sinke, C. B., & de Gelder, B. (2014). Affective scenes influence fear perception of individual body expressions. *Human brain mapping, 35*(2), 492-502.

Donderi, D. C. (2006). Visual complexity: a review. *Psychological bulletin, 132*(1), 73.

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., . . . Baas, J. M. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and anxiety, 32*(4), 239-253.

Dunsmoor, J. E., Kragel, P. A., Martin, A., & LaBar, K. S. (2013). Aversive learning modulates cortical representations of object categories. *Cerebral cortex, 24*(11), 2859-2872.

Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological psychology, 89*(2), 300-305.

Dunsmoor, J. E., & Paz, R. (2015). Fear generalization and anxiety: behavioral and neural mechanisms. *Biological psychiatry, 78*(5), 336-343.

Dunsmoor, J. E., Prince, S. E., Murty, V. P., Kragel, P. A., & LaBar, K. S. (2011). Neurobehavioral mechanisms of human fear generalization. *NeuroImage, 55*(4), 1878-1888.

Dunsmoor, J. E., White, A. J., & LaBar, K. S. (2011). Conceptual similarity promotes generalization of higher order fear learning. *Learning & Memory, 18*(3), 156-160.

Dymond, S., Dunsmoor, J. E., Vervliet, B., Roche, B., & Hermans, D. (2015). Fear generalization in humans: systematic review and implications for anxiety disorder research. *Behavior therapy, 46*(5), 561-582.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and brain sciences, 21*(4), 449-467.

Edmiston, E. K., McHugo, M., Dukic, M. S., Smith, S. D., Abou-Khalil, B., Eggers, E., & Zald, D. H. (2013). Enhanced visual cortical activation for emotional stimuli is preserved in patients with unilateral amygdala resection. *Journal of Neuroscience, 33*(27), 11023-11031.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental psychology and nonverbal behavior, 1*(1), 56-75.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598.

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain.

Estes, Z., Golonka, S., & Jones, L. L. (2011). Thematic thinking: The apprehension and consequences of thematic relations. In *Psychology of learning and motivation* (Vol. 54, pp. 249-294): Elsevier.

Etkin, A., Egner, T., & Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in cognitive sciences, 15*(2), 85-93.

Etkin, A., & Wager, T. D. (2007). Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. *American journal of Psychiatry, 164*(10), 1476-1488.

Fendt, M., & Fanselow, M. S. (1999). The neuroanatomical and neurochemical basis of conditioned fear. *Neuroscience & Biobehavioral Reviews, 23*(5), 743-760.

Fenton, A. A. (2007). Neuroscience. Where am I? *Science (New York, NY), 315*(5814), 947-949.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., & Turner, R. (1998). Event-related fMRI: characterizing differential responses. *Neuroimage, 7*(1), 30-40.

Fullana, M., Harrison, B., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Molecular psychiatry, 21*(4), 500-508.

Gaißert, N., Bülthoff, H. H., & Wallraven, C. (2011). Similarity and categorization: From vision to touch. *Acta psychologica, 138*(1), 219-230.

Gallo, D. A., Foster, K. T., & Johnson, E. L. (2009). Elevated false recollection of emotional pictures in young and older adults. *Psychology and aging, 24*(4), 981.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological science, 5*(3), 152-158.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist, 52*(1), 45.

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2021). The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature Human Behaviour*, 1-11.

Goldstone, R. L., Medin, D. L., & Halberstadt, J. (1997). Similarity in context. *Memory & Cognition, 25*(2), 237-255.

Golonka, S., & Estes, Z. (2009). Thematic relations affect similarity via commonalities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1454.

Grandjean, D., Sander, D., & Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and cognition, 17*(2), 484-495.

Gray, K. L., Adams, W. J., Hedger, N., Newton, K. E., & Garner, M. (2013). Faces and awareness: low-level, not emotional factors determine perceptual dominance. *Emotion, 13*(3), 537.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2014). Affordances Provide a Fundamental Categorization Principle for Visual Scenes. *arXiv preprint arXiv:1411.5340*.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron, 21*(1), 191-202.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife, 7*, e32962.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1714), 20160102.

Gross, C. T., & Canteras, N. S. (2012). The many paths to fear. *Nature Reviews Neuroscience, 13*(9), 651-658.

Guntupalli, J. S., Wheeler, K. G., & Gobbini, M. I. (2016). Disentangling the representation of identity from head view along the human face processing pathway. *Cerebral cortex, 27*(1), 46-53.

Guo, C. C., Gorno-Tempini, M. L., Gesierich, B., Henry, M., Trujillo, A., Shany-Ur, T., . . . Rankin, K. P. (2013). Anterior temporal lobe degeneration produces widespread network-driven dysfunction. *Brain, 136*(10), 2979-2991.

Halberstadt, J. B., & Niedenthal, P. M. (1997). Emotional state and the use of stimulus dimensions in judgment. *Journal of Personality and Social Psychology, 72*(5), 1017.

Halberstadt, J. B., Niedenthal, P. M., & Kushner, J. (1995). Resolution of lexical ambiguity by emotional state. *Psychological Science, 6*(5), 278-282.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., . . . Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72*(2), 404-416.

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods, 48*(2), 510-527.

Heinzel, A., Bermpohl, F., Niese, R., Pfennig, A., Pascual-Leone, A., Schlaug, G., & Northoff, G. (2005). How do we modulate our emotions? Parametric fMRI reveals cortical midline structures as regions specifically involved in the processing of emotional valences. *Cognitive Brain Research, 25*(1), 348-358.

Henson, R. N., & Mouchlianitis, E. (2007). Effect of spatial attention on stimulus-specific haemodynamic repetition effects. *NeuroImage, 35*(3), 1317-1329.

Highley, J. R., Walker, M. A., Esiri, M. M., Crow, T. J., & Harrison, P. J. (2002). Asymmetry of the uncinate fasciculus: a post-mortem study of normal subjects and patients with schizophrenia. *Cerebral cortex, 12*(11), 1218-1224.

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology, 55*(9), 1830.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior research methods, 45*(3), 718-730.

Horton, M. S., & Markman, E. M. (1980). Developmental differences in the acquisition of basic and superordinate categories. *Child development*, 708-719.

Hrybouski, S., Aghamohammadi-Sereshki, A., Madan, C. R., Shafer, A. T., Baron, C. A., Seres, P., . . . Malykhin, N. V. (2016). Amygdala subnuclei response and connectivity during emotional processing. *NeuroImage, 133*, 98-110.

Hsu, L.-K., Tseng, W.-S., Kang, L.-W., & Wang, Y.-C. F. (2013). *Seeing through the expression: Bridging the gap between expression and emotion recognition.* Paper presented at the 2013 IEEE International Conference on Multimedia and Expo (ICME).

Iordan, M. C., Ellis, C., Osherson, D., & Cohen, J. (2017). The Relative Contribution of Features and Dimensions to Semantic Similarity. *Journal of Vision, 17*(10), 1245-1245.

Iordan, M. C., Greene, M. R., Beck, D. M., & Fei-Fei, L. (2015). Basic level category structure emerges gradually across human ventral visual cortex. *Journal of Cognitive Neuroscience, 27*(7), 1427-1446.

Jabbi, M., Bastiaansen, J., & Keysers, C. (2008). A common anterior insula representation of disgust observation, experience and imagination shows divergent functional connectivity pathways. *PloS one, 3*(8), e2939.

Jackson, R. L., Hoffman, P., Pobric, G., & Lambon Ralph, M. A. (2015). The nature and neural correlates of semantic association versus conceptual similarity. *Cerebral cortex, 25*(11), 4319-4333.

Jefferies, E., Patterson, K., Jones, R. W., Ralph, L., & Matthew, A. (2009). Comprehension of concrete and abstract words in semantic dementia. *Neuropsychology, 23*(4), 492.

Julian, J. B., Ryan, J., Hamilton, R. H., & Epstein, R. A. (2016). The occipital place area is causally involved in representing environmental boundaries during navigation. *Current Biology, 26*(8), 1104-1109.

Kheirbek, M. A., Klemenhagen, K. C., Sahay, A., & Hen, R. (2012). Neurogenesis and generalization: a new approach to stratify and treat anxiety disorders. *Nature neuroscience, 15*(12), 1613-1620.

Kim, J., Shinkareva, S. V., & Wedell, D. H. (2017). Representations of modality-general valence for videos and music derived from fMRI data. *NeuroImage, 148*, 42-54.

King, M. L., Groen, I. I., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage, 197*, 368-382.

Koch, A., Alves, H., Krüger, T., & Unkelbach, C. (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(8), 1171.

Kohn, N., Eickhoff, S. B., Scheller, M., Laird, A. R., Fox, P. T., & Habel, U. (2014). Neural network of cognitive emotion regulation—an ALE meta-analysis and MACM analysis. *NeuroImage, 87*, 345-355.

Korn, C. W., Staib, M., Tzovara, A., Castegnetti, G., & Bach, D. R. (2017). A pupil size response model to assess fear learning. *Psychophysiology, 54*(3), 330-343.

Kragel, P. A., & LaBar, K. S. (2016). Decoding the nature of emotion in the brain. *Trends in cognitive sciences, 20*(6), 444-455.

Kragel, P. A., Reddan, M. C., LaBar, K. S., & Wager, T. D. (2019). Emotion schemas are embedded in the human visual system. *Science advances, 5*(7), eaaw4358.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, 103*(10), 3863-3868.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in psychology, 3*, 245.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008b). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*, 4.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Kuppens, P., Tuerlinckx, F., Russell, J. A., & Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological Bulletin, 139*(4), 917.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron, 20*(5), 937-945.

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1634), 20120392.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): affective ratings of pictures and instruction manual. University of Florida, Gainesville*. Retrieved from

Lange, I., Goossens, L., Michielse, S., Bakker, J., Lissek, S., Papalini, S., . . . Wichers, M. (2017). Behavioral pattern separation and its link to the neural mechanisms of fear generalization. *Social cognitive and affective neuroscience, 12*(11), 1720-1729.

Laufer, O., Israeli, D., & Paz, R. (2016). Behavioral and neural mechanisms of overgeneralization in anxiety. *Current Biology, 26*(6), 713-722.

Laufer, O., & Paz, R. (2012). Monetary loss alters perceptual thresholds and compromises future decisions via amygdala and prefrontal networks. *Journal of Neuroscience, 32*(18), 6304-6311.

Leal, S. L., Tighe, S. K., Jones, C. K., & Yassa, M. A. (2014). Pattern separation of emotional information in hippocampal dentate and CA3. *Hippocampus, 24*(9), 1146-1155.

Leal, S. L., Tighe, S. K., & Yassa, M. A. (2014). Asymmetric effects of emotion on mnemonic interference. *Neurobiology of learning and memory, 111*, 41-48.

Leal, S. L., & Yassa, M. A. (2018). Integrating new findings and examining clinical applications of pattern separation. *Nature neuroscience, 21*(2), 163.

Leclerc, C. M., & Kensinger, E. A. (2008). Effects of age on detection of emotional information. *Psychology and aging, 23*(1), 209.

LeDoux, J. (2007). The amygdala. *Current biology, 17*(20), R868-R874.

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

Leppänen, J. M., & Hietanen, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological research, 69*(1), 22-29.

Levine, S. M., Kumpf, M., Rupprecht, R., & Schwarzbach, J. V. (2021). Supracategorical fear information revealed by aversively conditioning multiple categories. *Cognitive Neuroscience, 12*(1), 28-39.

Levine, S. M., Wackerle, A., Rupprecht, R., & Schwarzbach, J. V. (2018). The neural representation of an individualized relational affective space. *Neuropsychologia, 120*, 35-42.

Liao, H.-I., Yoneya, M., Kidani, S., Kashino, M., & Furukawa, S. (2016). Human pupillary dilation response to deviant auditory stimuli: Effects of stimulus properties and voluntary attention. *Frontiers in Neuroscience, 10*, 43.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General, 130*(1), 3.

Lin, Y.-H., Young, I. M., Conner, A. K., Glenn, C. A., Chakraborty, A. R., Nix, C. E., . . . Hormovas, J. (2020). Anatomy and White Matter Connections of the Inferior Temporal Gyrus. *World Neurosurgery, 143*, e656-e666.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *Behavioral and brain sciences, 35*(3), 121-143.

Lissek, S., Bradford, D. E., Alvarez, R. P., Burton, P., Espensen-Sturges, T., Reynolds, R. C., & Grillon, C. (2014). Neural substrates of classically conditioned fear-generalization in humans: a parametric fMRI study. *Social cognitive and affective neuroscience, 9*(8), 1134-1142.

Lissek, S., Rabin, S., Heller, R. E., Lukenbaugh, D., Geraci, M., Pine, D. S., & Grillon, C. (2009). Overgeneralization of conditioned fear as a pathogenic marker of panic disorder. *American journal of Psychiatry, 167*(1), 47-55.

Liu, M., Liu, C. H., Zheng, S., Zhao, K., & Fu, X. (2021). Reexamining the neural network involved in perception of facial expression: A meta-analysis. *Neuroscience & Biobehavioral Reviews*.

Lundqvist, D., Flykt. A., & Öhman, A. . (1998). *The Karolinska Directed Emotional Faces - KDEF*.

Machajdik, J., & Hanbury, A. (2010). *Affective image classification using features inspired by psychology and art theory.* Paper presented at the Proceedings of the 18th ACM international conference on Multimedia.

Mack, M. L., Wong, A. C.-N., Gauthier, I., Tanaka, J. W., & Palmeri, T. J. (2009). Time course of visual object categorization: Fastest does not necessarily mean first. *Vision research, 49*(15), 1961-1968.

Madan, C. R., Bayer, J., Gamer, M., Lonsdorf, T. B., & Sommer, T. (2018). Visual complexity and affect: ratings reflect more than meets the eye. *Frontiers in psychology, 8*, 2368.

Maguire, P., Maguire, R., & Cater, A. W. (2010). The influence of interactional semantic patterns on the interpretation of noun–noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 288.

Malach, R., Reppas, J., Benson, R., Kwong, K., Jiang, H., Kennedy, W., . . . Tootell, R. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, 92*(18), 8135-8139.

Mäntylä, M., Adams, B., Destefanis, G., Graziotin, D., & Ortu, M. (2016). *Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?* Paper presented at the Proceedings of the 13th International Conference on Mining Software Repositories.

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior research methods, 46*(2), 596-610.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of memory and language, 32*(4), 517-535.

Marks, L. E. (1987). On cross-modal similarity: Perceiving temporal patterns by hearing, touch, and vision. *Perception & Psychophysics, 42*(3), 250-256.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature, 379*(6566), 649.

Matsumoto, D., & Ekman, P. (2008). Facial expression analysis. *Scholarpedia, 3*(5), 4237.

Mattar, M. G., & Talmi, D. (2019). Patterns of Neural Oscillations in Emotional Memory Discrimination. *Neuron, 102*(4), 715-717.

Mehrabian, A. (1980). *Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies*: Oelgeschlager, Gunn & Hain Cambridge, MA.

Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of insula function. *Brain structure and function, 214*(5-6), 655-667.

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review, 101*(2), 343.

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18-31.

Mondloch, C. J., Nelson, N. L., & Horner, M. (2013). Asymmetries of influence: Differential effects of body postures on perceptions of emotional facial expressions. *PloS one, 8*(9), e73605.

Moscovitch, D. A., Rowa, K., Paulitzki, J. R., Ierullo, M. D., Chiang, B., Antony, M. M., & McCabe, R. E. (2013). Self-portrayal concerns and their relation to safety behaviors and negative affect in social anxiety disorder. *Behaviour research and therapy, 51*(8), 476-486.

Motzkin, J. C., Philippi, C. L., Wolf, R. C., Baskaya, M. K., & Koenigs, M. (2015). Ventromedial prefrontal cortex is critical for the regulation of amygdala activity in humans. *Biological psychiatry, 77*(3), 276-284.

Murphy, F. C., Nimmo-Smith, I., & Lawrence, A. D. (2003). Functional neuroanatomy of emotions: a meta-analysis. *Cognitive, affective, & behavioral neuroscience, 3*(3), 207-233.

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(1), 70.

Nestor, A., Plaut, D. C., & Behrmann, M. (2016). Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences, 113*(2), 416-421.

Nestor, P. J., Fryer, T. D., & Hodges, J. R. (2006). Declarative memory impairments in Alzheimer's disease and semantic dementia. *NeuroImage, 30*(3), 1010-1020.

Neyens, V., Bruffaerts, R., Liuzzi, A. G., Kalfas, I., Peeters, R., Keuleers, E., . . . Dupont, P. (2017). Representation of semantic similarity in the left intraparietal sulcus: functional magnetic resonance imaging evidence. *Frontiers in human neuroscience, 11*, 402.

Niedenthal, P. M., Halberstadt, J. B., & Innes-Ker, Å. H. (1999). Emotional response categorization. *Psychological review, 106*(2), 337.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS computational biology, 10*(4), e1003553.

Ohira, H., Nomura, M., Ichikawa, N., Isowa, T., Iidaka, T., Sato, A., . . . Yamada, J. (2006). Association of neural and physiological responses during voluntary emotion suppression. *NeuroImage, 29*(3), 721-733.

Öhman, A. (2009). Of snakes and faces: An evolutionary perspective on the psychology of fear. *Scandinavian journal of psychology, 50*(6), 543-552.

Olson, I. R., McCoy, D., Klobusicky, E., & Ross, L. A. (2013). Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social cognitive and affective neuroscience, 8*(2), 123-133.

Onat, S., & Büchel, C. (2015). The neuronal basis of fear generalization in humans. *Nature neuroscience, 18*(12), 1811-1818.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin, 49*(3), 197.

Palermo, R., & Coltheart, M. (2004). Photographs of facial expression: Accuracy, response times, and ratings of intensity. *Behavior Research Methods, Instruments, & Computers, 36*(4), 634-638.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*(12), 976.

Paulus, M. P., & Stein, M. B. (2006). An insular view of anxiety. *Biological psychiatry, 60*(4), 383-387.

Pavlov, I. P. (1927). Conditioned reflexes (translated by GV Anrep). *London: Oxford*.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage, 16*(2), 331-348.

Plichta, M. M., Grimm, O., Morgen, K., Mier, D., Sauer, C., Haddad, L., . . . Schwarz, A. J. (2014). Amygdala habituation: a reliable fMRI phenotype. *NeuroImage, 103*, 383-390.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist, 89*(4), 344-350.

Plutchik, R. E., & Conte, H. R. (1997). *Circumplex models of personality and emotions*: American Psychological Association.

Pobric, G., Jefferies, E., & Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proceedings of the National Academy of Sciences, 104*(50), 20137-20141.

Pobric, G., Lambon Ralph, M. A., & Zahn, R. (2016). Hemispheric specialization within the superior anterior temporal cortex for social and nonsocial concepts. *Journal of Cognitive Neuroscience, 28*(3), 351-360.

Pochedly, J. T., Widen, S. C., & Russell, J. A. (2012). What emotion does the "facial expression of disgust" express? *Emotion, 12*(6), 1315.

Pollak, S. D., Cicchetti, D., Hornung, K., & Reed, A. (2000). Recognizing emotion in faces: developmental effects of child abuse and neglect. *Developmental psychology, 36*(5), 679.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology, 17*(3), 715-734.

Prince, J. S., & Konkle, T. (2020). Computational evidence for integrated rather than specialized feature tuning in category-selective regions. *Journal of Vision, 20*(11), 1577-1577.

Puccetti, N. A., Schaefer, S. M., Van Reekum, C. M., Ong, A. D., Almeida, D. M., Ryff, C. D., . . . Heller, A. S. (2021). Linking amygdala persistence to real-world emotional experience and psychological well-being. *Journal of Neuroscience, 41*(16), 3721-3730.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience, 18*(1), 42.

Ralph, M. A. L., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences, 107*(6), 2717-2722.

Ralph, M. L., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain, 130*(4), 1127-1137.

Resnik, J., Sobel, N., & Paz, R. (2011). Auditory aversive learning increases discrimination thresholds. *Nature neuroscience, 14*(6), 791-796.

Riberto, M., Pobric, G., & Talmi, D. (2019). The emotional facet of subjective and neural indices of similarity. *Brain topography, 32*(6), 956-964.

Riberto, M., Pobric, G., & Talmi, D. (2020). A Response to 'Investigating Emotional Similarity: A Comment on Riberto, Pobric and Talmi (2019)'. *Brain topography, 33*(3), 288-288.

Riberto, M., Talmi, D., & Pobric, G. (2021). Symmetry in emotional and visual similarity between neutral and negative faces. *Symmetry, 13*(11), 2091.

Riddoch, M. J., Humphreys, G. W., Coltheart, M., & Funnell, E. (1988). Semantic systems or system? Neuropsychological evidence re-examined. *Cognitive Neuropsychology, 5*(1), 3-25.

Riegel, M., Żurawski, Ł., Wierzba, M., Moslehi, A., Klocek, Ł., Horvat, M., . . . Marchewka, A. (2016). Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE). *Behavior research methods, 48*(2), 600-612.

Roberts, J. S., & Wedell, D. H. (1994). Context effects on similarity judgments of multidimensional stimuli: Inferring the structure of the emotion space. *Journal of Experimental Social Psychology, 30*(1), 1-38.

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General, 136*(3), 451.

Rogers, T. T., Ralph, L., Matthew, A., Garrard, P., Bozeat, S., McClelland, J. L., . . . Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological review, 111*(1), 205.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology, 8*(3), 382-439.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology, 39*(6), 1161.

Russell, J. A., & Bullock, M. (1985). Multidimensional scaling of emotional facial expressions: similarity from preschoolers to adults. *Journal of Personality and Social Psychology, 48*(5), 1290.

Russell, J. A., & Pratt, G. (1980). A description of the affective quality attributed to environments. *Journal of personality and social psychology, 38*(2), 311.

Saarimäki, H., Ejtehadian, L. F., Glerean, E., Jääskeläinen, I. P., Vuilleumier, P., Sams, M., & Nummenmaa, L. (2018). Distributed affective space represents multiple emotion categories across the human brain. *Social cognitive and affective neuroscience, 13*(5), 471-482.

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision, 10*(5), 11-11.

Sakaki, M., Niki, K., & Mather, M. (2012). Beyond arousal and valence: The importance of the biological versus social relevance of emotional stimuli. *Cognitive, Affective, & Behavioral Neuroscience, 12*(1), 115-139.

Sander, D., Grandjean, D., & Scherer, K. R. (2018). An appraisal-driven componential approach to the emotional brain. *Emotion Review, 10*(3), 219-231.

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience, 30*(31), 10460-10464.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of experimental psychology, 44*(4), 229.

Schwartz, M. F., Kimberg, D. Y., Walker, G. M., Brecher, A., Faseyitan, O. K., Dell, G. S., . . . Coslett, H. B. (2011). Neuroanatomical dissociation for taxonomic and thematic knowledge in the human brain. *Proceedings of the National Academy of Sciences, 108*(20), 8520-8524.

Segal, S. K., Stark, S. M., Kattan, D., Stark, C. E., & Yassa, M. A. (2012). Norepinephrine-mediated emotional arousal facilitates subsequent pattern separation. *Neurobiology of learning and memory, 97*(4), 465-469.

Sehlmeyer, C., Schöning, S., Zwitserlood, P., Pfleiderer, B., Kircher, T., Arolt, V., & Konrad, C. (2009). Human fear conditioning and extinction in neuroimaging: a systematic review. *PloS one, 4*(6), e5865.

Shalev, L., Paz, R., & Avidan, G. (2018). Visual aversive learning compromises sensory discrimination. *Journal of Neuroscience, 38*(11), 2766-2779.

Sharma, P., Esengönül, M., Khanal, S. R., Khanal, T. T., Filipe, V., & Reis, M. J. (2018). *Student concentration evaluation index in an e-learning context using facial emotion analysis.* Paper presented at the International Conference on Technology and Innovation in Learning, Teaching and Education.

Shinkareva, S. V., Wang, J., & Wedell, D. H. (2013). Examining similarity structure: multidimensional scaling and related approaches in neuroimaging. *Computational and mathematical methods in medicine, 2013*.

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition, 108*(3), 781-795.

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Josa a, 4*(3), 519-524.

Sison, J. A. G., & Mather, M. (2007). Does remembering emotional items impair recall of same-emotion items? *Psychonomic bulletin & review, 14*(2), 282-287.

Sloutsky, V. M., & Napolitano, A. C. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child development, 74*(3), 822-833.

Soto, J. A., Armenta, B. E., Perez, C. R., Zamboanga, B. L., Umaña-Taylor, A. J., Lee, R. M., . . . Whitbourne, S. K. (2012). Strength in numbers? Cognitive reappraisal tendencies and psychological functioning among Latinos in the context of oppression. *Cultural Diversity and Ethnic Minority Psychology, 18*(4), 384.

Spielberger, C. D., Gorsuch, R., Lushene, R., Vagg, P., & Jacobs, G. (1983). State-trait anxiety inventory. Palo Alto. *CA: Mind Garden*.

Starita, F., Kroes, M. C., Davachi, L., Phelps, E. A., & Dunsmoor, J. E. (2019). Threat learning promotes generalization of episodic memory. *Journal of Experimental Psychology: General, 148*(8), 1426.

Struyf, D., Zaman, J., Hermans, D., & Vervliet, B. (2017). Gradients of fear: How perception influences fear generalization. *Behaviour research and therapy, 93*, 116-122.

Talmi, D. (2013). Enhanced emotional memory: Cognitive and neural mechanisms. *Current Directions in Psychological Science, 22*(6), 430-436.

Talmi, D., Luk, B. T., McGarry, L. M., & Moscovitch, M. (2007). The contribution of relatedness and distinctiveness to emotionally-enhanced memory. *Journal of Memory and Language, 56*(4), 555-574.

Talmi, D., & McGarry, L. M. (2012). Accounting for immediate emotional memory enhancement. *Journal of Memory and Language, 66*(1), 93-108.

Talmi, D., & Moscovitch, M. (2004). Can semantic relatedness explain the enhancement of memory for emotional words? *Memory & Cognition, 32*(5), 742-751.

Taylor, K. I., Devereux, B. J., Acres, K., Randall, B., & Tyler, L. K. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition, 122*(3), 363-374.

Todd, R. M., Miskovic, V., Chikazoe, J., & Anderson, A. K. (2020). Emotional objectivity: Neural representations of emotions and their interaction with cognition. *Annual review of psychology, 71*, 25-48.

Todd, R. M., Schmitz, T. W., Susskind, J., & Anderson, A. K. (2013). Shared neural substrates of emotionally enhanced perceptual and mnemonic vividness. *Frontiers in behavioral neuroscience, 7*, 40.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry research, 168*(3), 242-249.

Tseng, A., Bansal, R., Liu, J., Gerber, A. J., Goh, S., Posner, J., . . . Russell, J. A. (2014). Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. *Journal of autism and developmental disorders, 44*(6), 1332-1346.

Tversky, A. (1977). Features of similarity. *Psychological review, 84*(4), 327.

üge Çarıkçı, M., & Özen, F. (2012). A face recognition system based on eigenfaces method. *Procedia Technology, 1*, 118-123.

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology, 69*(10), 1996-2019.

van Tilburg, W. A., & Igou, E. R. (2017). Boredom begs to differ: Differentiation from other negative emotions. *Emotion, 17*(2), 309.

Visser, M., Jefferies, E., Embleton, K. V., & Lambon Ralph, M. A. (2012). Both the middle temporal gyrus and the ventral anterior temporal area are crucial for multimodal semantic processing: distortion-corrected fMRI evidence for a double gradient of information convergence in the temporal lobes. *Journal of Cognitive Neuroscience, 24*(8), 1766-1778.

Visser, R. M., de Haan, M. I., Beemsterboer, T., Haver, P., Kindt, M., & Scholte, H. S. (2016). Quantifying learning-dependent changes in the brain: Single-trial multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology, 53*(8), 1117-1127.

Visser, R. M., Scholte, H. S., Beemsterboer, T., & Kindt, M. (2013). Neural pattern similarity predicts long-term fear memory. *Nature neuroscience, 16*(4), 388-390.

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. *Journal of Neuroscience, 31*(33), 12021-12028.

Von Der Heide, R. J., Skipper, L. M., Klobusicky, E., & Olson, I. R. (2013). Dissecting the uncinate fasciculus: disorders, controversies and a hypothesis. *Brain, 136*(6), 1692-1707.

Vytal, K., & Hamann, S. (2010). Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. *Journal of Cognitive Neuroscience, 22*(12), 2864-2885.

Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., & Barrett, L. F. (2015). A Bayesian model of category-specific emotional brain responses. *PLoS computational biology, 11*(4), e1004066.

Wagner, I. C., Rütgen, M., & Lamm, C. (2020). Pattern similarity and connectivity of hippocampal-neocortical regions support empathy for pain. *Social cognitive and affective neuroscience, 15*(3), 273-284.

Wang, Y., Collins, J. A., Koski, J., Nugiel, T., Metoki, A., & Olson, I. R. (2017). Dynamic neural architecture for social knowledge retrieval. *Proceedings of the National Academy of Sciences, 114*(16), E3305-E3314.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin, 98*(2), 219.

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PloS one, 12*(5), e0177239.

Wheelock, M. D., Sreenivasan, K. R., Wood, K. H., Ver Hoef, L., Deshpande, G., & Knight, D. (2014). Threat-related learning relies on distinct dorsal prefrontal cortex network connectivity. *NeuroImage, 102*, 904-912.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V., & Rizzolatti, G. (2003). Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron, 40*(3), 655-664.

Wiemer, J., Rauner, M. M., Stegmann, Y., & Pauli, P. (2021). Reappraising fear: is up-regulation more efficient than down-regulation? *Motivation and Emotion, 45*(2), 221-234.

Wierzba, M., Riegel, M., Pucz, A., Leśniewska, Z., Dragan, W. Ł., Gola, M., . . . Marchewka, A. (2015). Erotic subset for the Nencki Affective Picture System (NAPS ERO): cross-sexual comparison study. *Frontiers in psychology, 6*, 1336.

Wisniewski, E. J., & Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive psychology, 39*(3-4), 208-238.

Xiao, X., Dong, Q., Chen, C., & Xue, G. (2016). Neural pattern similarity underlies the mnemonic advantages for living words. *Cortex, 79*, 99-111.

Xu, Y., Wang, X., Wang, X., Men, W., Gao, J.-H., & Bi, Y. (2018). Doctor, teacher, and stethoscope: neural representation of different types of semantic relations. *Journal of Neuroscience, 38*(13), 3303-3317.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods, 8*(8), 665.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., . . . Zhang, X. (2016). *Building Chinese affective resources in valence-arousal dimensions.* Paper presented at the Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Yuan, J., Mcdonough, S., You, Q., & Luo, J. (2013). *Sentribute: image sentiment analysis from a mid-level perspective.* Paper presented at the Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining.

Yuen, K., Johnston, S., Martino, F., Sorger, B., Formisano, E., Linden, D., & Goebel, R. (2012). Pattern classification predicts individuals' responses to affective stimuli. *Translational Neuroscience, 3*(3), 278-287.

Zahn, R., Moll, J., Iyengar, V., Huey, E. D., Tierney, M., Krueger, F., & Grafman, J. (2009). Social conceptual impairments in frontotemporal lobar degeneration with right anterior temporal hypometabolism. *Brain, 132*(3), 604-616.

Zahn, R., Moll, J., Krueger, F., Huey, E. D., Garrido, G., & Grafman, J. (2007). Social concepts are represented in the superior anterior temporal cortex. *Proceedings of the National Academy of Sciences, 104*(15), 6430-6435.

Zeithamova, D., Gelman, B. D., Frank, L., & Preston, A. R. (2018). Abstract representation of prospective reward in the hippocampus. *Journal of Neuroscience, 38*(47), 10093-10101.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology, 43*(1), 111.

Zhao, S., Ding, G., Huang, Q., Chua, T.-S., Schuller, B. W., & Keutzer, K. (2018). *Affective Image Content Analysis: A Comprehensive Survey.* Paper presented at the IJCAI.

Zheng, J., Stevenson, R. F., Mander, B. A., Mnatsakanyan, L., Hsu, F. P., Vadera, S., . . . Lin, J. J. (2019). Multiplexing of theta and alpha rhythms in the amygdala-hippocampal circuit supports pattern separation of emotional information. *Neuron, 102*(4), 887-898. e885.