Preprint version of paper forthcoming in: Eugen Fischer and Mark Curtis (eds.). *Methodological Advances in Experimental Philosophy*. London: Bloomsbury, 2019.

Using fMRI in experimental philosophy: Exploring the prospects

Rodrigo Díaz

Abstract: This chapter analyses the prospects of using neuroimaging methods, in particular functional magnetic resonance imaging (fMRI), for philosophical purposes. To do so, it will use two case studies from the field of emotion research: Greene et al. (2001) used fMRI to uncover the mental processes underlying moral intuitions, while Lindquist et al. (2012) used fMRI to inform the debate around the nature of a specific mental process, namely, emotion. These studies illustrate two main approaches in cognitive neuroscience: Reverse inference and ontology testing, respectively. With regards to Greene et al.'s study, the use of Neurosynth (Yarkoni 2011) will show that the available formulations of reverse inference, although viable a priori, seem to be of limited use in practice. On the other hand, the discussion of Lindquist et al.'s study will present the so far neglected potential of ontology-testing approaches to inform philosophical questions.

Keywords: fMRI, neuroimaging, reverse inference, cognitive ontology, emotion, experimental philosophy.

1. Introduction

Experimental philosophy is an interdisciplinary approach that consists in investigating traditional philosophical questions using experimental methods from the social sciences (Knobe and Nichols 2017). After a first period in which experimental philosophers basically relied on the use of questionnaires, the field has grown to adopt a wide variety of methods from other disciplines such as linguistics or neuroscience (see other chapters in this volume for several examples). The present chapter will analyse the possibilities of

neuroimaging methods, fMRI in particular, within the practice of experimental philosophy.

1.1. fMRI

Functional Magnetic Resonance Imaging (fMRI) is probably the most used neuroimaging method in cognitive neuroscience. Most research in cognitive neuroscience is committed to the goal of localizing mental processes in the brain[1], and fMRI serves this purpose well. Researchers use fMRI to measure activity in the brain of the participants in their experiments. Used while participants perform cognitive tasks, it provides information about the neural correlates of those tasks, which is used to make links between brain structure and mental function. The localization of the brain activity found in an experiment is usually represented through colouring superimposed on brain pictures. The resulting 'neuroimages' are not like photographs, however, and there is a considerable number of inferential steps from the acquisition of raw fMRI data to the attainment of brain activity maps (Roskies 2008).

The details of fMRI technology are complex and I can only explain a few key points here (Logothetis 2008; Klein 2010; Parens and Johnston 2014). First, fMRI is an *indirect* measure of brain activity. The fMRI signal measures changes in the oxygenation of blood in the research subject's brain. These changes are detected taking advantage of the different magnetic properties of deoxygenated haemoglobin in comparison to oxygenated haemoglobin in cerebral blood flow. As brain activity consumes oxygen, it is inferred that more deoxygenated haemoglobin in a particular brain area equals activity in that area. Second, the signal has limitations in terms of *temporal* and *spatial* resolution: changes in blood oxygenation are slower than brain activity, and the signal is recorded in volume units called *voxels* that include millions of neurons. Finally, and most importantly, fMRI results are *qualitative*. The final neuroimage is the result of complex mathematical and statistical processing of the signal, and it does not map the intensity of the signal, but the statistical significance of differences in the signal between experimental conditions. However, despite these limitations, fMRI is generally assumed to be suitable for the purpose of associating mental processes with their neural substrates, mapping mental function to brain structure.

[1] But see section 4.1

1.2. Experimental philosophy

To evaluate the prospects of using fMRI within experimental philosophy, it is necessary to explain the ways in which this movement employs experimental methods to address philosophical questions. For this purpose, it is useful to differentiate two different projects within the practice of experimental philosophy based on their relation with the analytical philosophy tradition (Rose and Danks 2013; Fischer and Collins 2015; Knobe and Nichols 2017): The *intuitions project*, and the *psychology project*.

The intuitions project builds on the use – widespread in analytic philosophy – of the method of cases: using hypothetical scenarios to trigger intuitions about a particular topic (e.g. knowledge, moral responsibility, free will). But instead of relying on 'armchair reflection', experimental philosophers use surveys and questionnaires to investigate people's intuitions, as well as the factors that underlie those intuitions: psychological mechanisms, cultural background, etc. The ultimate goal behind this investigation differentiates between two programs within the intuitions project: the positive program, and the negative program. Work in the positive program collects data to make progress on the understanding of the concept at hand, while work on the negative program tries to raise questions about the validity of the method of cases by showing that intuitions depend on unreliable mental processes or are affected by putatively irrelevant factors (Machery 2017).

The psychology project, in contrast to the intuitions project, is not primarily concerned with the study of intuitions and does not necessarily engage with the method of cases. Instead of collecting data on people's thoughts and feelings to inform some further philosophical issue, the issue is people's thoughts and feelings themselves. Work on this project blurs the boundaries between disciplines, as most questions involved are both questions of philosophy and questions of psychology. For example, the question of how people come to attribute mental states to others is of interest to both psychologists (mentalizing processes) and philosophers (the problem of other minds).

Within the intuitions project, using fMRI can help to uncover the mental processes underlying people's intuitions about hypothetical cases. An example of this approach is Greene et al.'s 2001 study on the impact of emotion on responses to moral dilemmas. In

this study, fMRI is used to identify the mental processes underlying different moral intuitions to assess their epistemological warrant.

Within the psychology project, fMRI data can *also* be used in a different way, to inform questions about cognitive architecture. An example of this approach is Lindquist et al.'s (2012) meta-analysis of studies investigating the brain basis of emotion. In this study, fMRI data is used to elucidate the nature of emotion and examine whether emotions are 'basic' or 'constructed'.

In the following, I will use Greene et al. (2001) and Lindquist et al. (2012) as case studies to present and discuss the above two ways in which fMRI could be used to inform philosophical work. These two influential papers[2] are selected because while both use fMRI to study the same mental process, emotion, they use different approaches. By looking at fMRI-recorded patterns of brain activity, Greene et al. aim to infer the *engagement* of emotion, while Lindquist et al. use fMRI to infer the *nature* of emotions. The former strategy is an instance of reverse inference (from brain activity to *what* the mind is doing), the latter constitutes an example of ontology testing (from brain activity to *how* the mind is organized). The problems associated with reverse inference, in particular the lack of one-to-one mappings between mental function and brain structure, is one of the aspects that motivate the use of ontology testing approaches. The discussion of Greene et al. (2001) and Lindquist et al. (2012) will provide a concrete illustration of how information about the neural substrates of a particular mental process (e.g. emotion) cannot support reverse inferences but can tell us something about the nature of that mental process. Furthermore, each study exemplifies one of the most used methods within each strategy: subtraction within reverse inference, and meta-analysis within ontology testing (see table 1). The discussion of these two studies will serve to argue against reverse inference and in favour of ontology testing approaches within experimental philosophy.

2 According to Google Scholar, Greene et al. (2001) has been cited 3656 times, while Lindquist et al. (2012) has been cited 1067 times (date of the search: 14.03.2018).

Study	Project	Strategy	Method
Greene et al. (2001)	Intuitions project	Reverse inference	Subtraction
Lindquist et al. (2012)	Psychology project	Ontology testing	Meta-analysis

Table 1. Project, strategy, and method of the studies discussed. Note that there is no one-to-one correspondence between method, strategy and project. For example, subtraction is not only used for reverse inference, and reverse inference is not only used within experimental philosophy's intuitions project.

2. Greene et al. (2001): Intuitions Project and Reverse inference

2.1. Research context

Greene et al.'s 2001 study uses a *subtractive design* common in cognitive neuroscience. This design usually involves two tasks, which differ in one specific task component. For example: task T_1 (hear a 'beep' sound), consisting in one component C_1 (auditory), and task T_2 (tap a finger when hear a 'beep' sound) involving components C_1 (auditory) and C_2 (motor). The design tries to *isolate* the differing component (C_2), which is the target of investigation. Neuroimaging is used to record participants' brain activity while they perform each task, and the recorded pattern of brain activity in the contrast task (T_1) is *subtracted* from the pattern of brain activity in the task that involves the component of interest (T_2). This subtraction consists on performing a statistical test for each voxel to see if it shows a significant difference in activity between the two tasks. The result is the identification of differential activity in certain areas of the brain, which is taken to be the neural correlate of the investigated task component (C_2). This design can be used for two different goals: to identify the neural basis of a specific mental process, or to identify the mental process(es) involved in a task.

First, a subtractive design can aim at identifying the neural basis of a specific mental process. If the isolated task component is assumed to correspond to the involvement of a specific mental process, the differential pattern of activity is interpreted as the neural basis of that mental process. For example, on this approach, the differential brain activity

between listening to emotionally un-evocative music and listening to emotionally evocative music is interpreted as the neural basis of emotion (Mitterschiffthaler et al. 2007). This strategy is known as forward inference, and it serves to establish 'structure-function links', links between localized brain activity and mental processes.

Second, subtractive designs can be used to identify the mental process(es) involved in a task. In some cases, the mental process(es) associated with the isolated task component are unknown, and subtraction is used to identify them. For example, contrasting an evaluative judgment task with a factual judgment task can serve to investigate the mental processes underlying evaluation (Moll et al. 2001). In these cases, the differential pattern of brain activity between one task and the other is interpreted in the light of previously established structure-function links. The logic is the following: (1) In task T, activation is found in region R; (2) In previous studies, activity in region R has been associated with mental process M; therefore (3) T recruits M. This is known as reverse inference: inferring the involvement of a particular mental process by looking at the localization of brain activity. Greene et al.'s study follows this strategy: It uses subtraction to study the mental processes underlying moral intuitions.

2.2. Greene et al.'s study

Participants in Greene et al.'s study were presented with a battery of dilemmas while their brain activity was recorded using an fMRI scanner. There were three different types of dilemmas: personal moral dilemmas, impersonal moral dilemmas, and non-moral dilemmas (which served as a control condition). Non-moral dilemmas involve decisions such as which of two coupons to use at a store. The paradigmatic example of an impersonal moral dilemma is the switch dilemma: A runaway trolley is about to run over five persons, and you have to decide whether to hit a switch that turns the trolley onto a different set of tracks in which it will only kill one person instead of five. Its personal counterpart is the footbridge dilemma: A runaway trolley is about to run over five persons, and you have to decide whether to push a fat person onto the tracks in front of the trolley in order to stop it, saving the other five people but killing the one that you push.

Both personal and impersonal dilemmas involve the same trade-off of lives: sacrifice one person to save five; but they differ in the kind of action required: pulling a switch versus pushing the person to sacrifice. Performing these actions is interpreted as an 'utilitarian'

response (choosing the better outcome), while inaction is interpreted as a 'deontological' response (not violating someone's rights). It is a consistent finding that, despite both types of dilemmas involving the same trade-off of lives, most people give utilitarian responses in impersonal dilemmas but deontological responses in personal dilemmas. Greene and colleagues hypothesize that this asymmetric pattern of responses is due to the higher emotional salience of personal dilemmas in comparison to impersonal ones. The deontological intuitions triggered by personal dilemmas would be grounded in people's emotional reactions, while utilitarian responses would be triggered in more 'cognitive' settings. They used fMRI to test this hypothesis via reverse inference.

When comparing the fMRI data relative to each category of dilemmas, Greene et al. found that personal moral dilemmas, in comparison to impersonal and non-moral dilemmas, are more likely to trigger activity in brain areas associated with emotion (Brodmann's Areas 9-10, 31 and 39). Conversely, personal dilemmas are less likely to trigger activity in areas associated with working memory (Brodmann's Areas 46 and 7/40).

2.3. Discussion

The results seem to support Greene and colleagues' hypothesis: personal dilemmas differ from impersonal ones in that they generate emotional reactions in participants, what would in turn explain the prevalence of 'deontological' responses to this type of dilemma. This evidence is supposed to undermine the epistemological warrant of deontological intuitions (Greene 2014). Moral dilemmas are unfamiliar situations, and in unfamiliar situations (e.g. driving a car for the first time) we shouldn't rely on 'automatic' but rather 'controlled' processes. As the results suggest that deontological responses to moral dilemmas are associated with 'automatic' emotional reactions, this would undermine deontological intuitions. However, it is important to take a closer look at the evidence that supports Greene et al.'s reverse inference from the fMRI data to the engagement of emotion. The inference has the following structure:

(1) When responding to personal moral dilemmas, activity in areas 9-10, 31 and 39 is found.
(2) Activity in areas 9-10, 31 and 39 has been associated with emotion.
(3) Therefore, responding to personal moral dilemmas involves emotion.

In order for (3) to follow from (1) and (2), there must be a one-to-one correspondence between brain structure and cognitive function (Poldrack 2006): emotion should consistently recruit activity in areas 9-10, 31 and 39; and activity in areas 9-10, 31 and 39 should not be consistently recruited by other mental processes. Deducing the engagement of a mental process from activity in a particular brain area is justified *only if* no other mental processes have been associated with activity in that area. However, as Klein (2011) has already noted, none of the areas reported by Greene and colleagues are *selective* for emotion. Instead, they are *pluripotent*: they are involved in the realization of multiple different mental processes. Thus, we cannot know whether activity in those areas means that emotion or one of the other processes that have been associated with those areas are engaged. Activity in areas 9-10, 31, and 39 is *not sufficient* to infer emotion engagement.

However, selectivity might be an excessive requirement for reverse inference. Although activity in areas 9-10, 31 and 39 is not *selective* for emotion, it might be that activity in those areas is *preferentially* associated with emotion. That is, when there is activity in those areas, the probability that emotion is being engaged is higher than the probability that other mental processes are being engaged. By reformulating reverse inference in probabilistic terms, it is possible to use fMRI data to provide some support for cognitive hypotheses (Poldrack 2006). To calculate the probability of emotion engagement given activation in areas 9-10, 31, and 39, we need to know how many times those regions were active when emotion was engaged, and how many times those regions were active when emotion was not engaged. Although the latter information is difficult to obtain, fMRI databases can be used to estimate the probability of a mental process being engaged given activation in a particular brain area.

Neurosynth (Yarkoni et al. 2011) is an automated database of fMRI studies that can help determine the *selectivity* and *consistency* of structure-function links. Neurosynth uses text-mining techniques to provide information about the probability of finding a term (e.g. "emotion", "memory", "attention"…) in the abstract of a paper when activation in a specific brain area is reported in that paper (reverse inference / selectivity) and, conversely, the probability of finding activation across different brain areas given the presence of a specific term (forward inference / consistency). A search in Neurosynth[3]

[3] The Talairach coordinates for peak activations reported by Greene et al. were transformed to MNI space (Lacadie et al. 2008) to enable the search in Neurosynth (date of the search: 14.03.2018). Full results can be found in the

revealed that, within the areas that Greene and colleagues reported, only area 9-10 is consistently associated with emotion ($z = 6.93$), and it is thus the only candidate to provide evidence for the engagement of emotion. There was a high probability of finding the term 'emotional' (.71) or 'affective' (.68) given activity in this area.[4] However, this probability was as high as the probability of finding other terms such as 'mentalizing' (.85), 'theory of mind' (.83), 'social cognition' (.83), 'self-referential' (.79), 'inferences' (.83), 'evaluation' (.76), 'intentions' (.82), 'belief' (.84), 'autobiographical' (.80), 'default mode' (.71), 'judgments' (.73), 'moral' (.82), or 'economic' (.81) (see note 5 for a complete list). The data suggests that the selectivity of area 9-10 for emotion is low. Thus, the presence of activity in this area is not able to provide strong support to Greene et al.'s hypothesis.

Nevertheless, one could argue that the possibility of mentalizing, theory of mind, social cognition, etc. being involved should not be taken into account. The only mental processes that should be considered are those relevant in the case at hand. Although this possibility is not systematically investigated by Greene and colleagues, some have argued that their study could help deciding whether deontological responses to personal moral dilemmas are generated by (H1) emotional reactions to the vignette or (H2) the application of an abstract moral rule, e.g. the doctrine of double effect (Del Pinal and Nathan 2013; Machery 2013). In this case, the only mental processes to take into account would be (1) emotion and (2) rule application. Although fMRI evidence cannot provide strong support to a cognitive hypothesis in isolation, it could help selecting between two competing hypotheses. Our search in Neurosynth for areas 9-10 and 31 revealed that the use of the term 'rule' was not consistently associated with activity in any of these areas (see note 5). Thus, in this comparative framework, Greene et al.'s results would support H1 over H2, as activity in area 9-10 is more likely to be found when reacting emotionally than when applying an abstract rule.

However, the hypothesis that participants are applying an abstract rule when responding to personal moral dilemmas is compatible with those subjects experiencing emotion (Huebner et al. 2009; Mole and Klein 2010). Emotion and rule application can occur

following links: BA 9-10 (http://neurosynth.org/locations/2_56_20_6/) BA 31 (http://neurosynth.org/locations/-6_-58_38_6/). The coordinates for BA 39 are not provided in Greene et al.'s paper.

[4] Some terms related to emotion such as 'empathy' (.78) or 'unpleasant' (.79) could arguably support Greene et al.'s hypothesis too, while others like 'positive' (.67) don't.

alongside each other. Thus, even if it is the case that activity in area 9-10 is due to emotional engagement, we should not prefer H1 over H2. In order for the neuroimaging data to help us select between these two competing hypotheses, we need strong structure-function links for both of them. This is necessary to contrast the predictions of H1 and H2 one against the other. H1 posits that personal moral dilemmas, in contrast to impersonal ones, engage emotional processes. Thus, it predicts differential activity in areas associated with emotion for personal dilemmas. Conversely, H2 posits that responding to both personal and impersonal moral dilemmas involve the application of a rule: a deontological rule in the former, and a consequentialist rule in the latter. Thus, it predicts activity in areas associated with rule application for both types of dilemmas.

We might say that the neuroimaging data was in line with H1 predictions[5], but was it *against* H2 predictions? For this, we need to know the link between rule application and localized brain activity. Although emotions have been the target of much neuroimaging work, this is not the case for rule application.[6] In Greene et al.'s study, rule application processes would putatively be supported by areas 46 and 7/40, which showed diminished activity in personal dilemmas in comparison to impersonal ones. This pattern of activity would favour H1 over H2. However, the consistency and selectivity of these areas for rule application is important at this point. If the link is weak, a proponent of H2 could argue that other areas which didn't show differential activity, and not areas 46 and 7/40, are the ones supporting rule application. A search in Neurosynth[7] revealed that activity in areas 46 and 7/40 was not consistently associated with the term 'rule'. Thus, the evidence does not undermine H2, and H1 should not be preferred over it.

The use of Neurosynth and Greene et al.'s study has shown the problems that the different formulations of reverse inference face when they are to be applied. It is important to note that the problems exposed in this section are not exclusive to Greene et al.'s study, but of reverse inference in general. Lack of selectivity is not an anomaly of some brain areas,

[5] One could still question why differential activity was only found in 9-10 but not in other areas that are also consistently associated with emotion.

[6] Using Neurosynth, 790 studies were found to be associated with the term 'emotion' (http://neurosynth.org/analyses/terms/emotion/). There were no results for 'rule application', and only 141 studies associated with the term 'rule' (http://neurosynth.org/analyses/terms/rule/)

[7] The Talairach coordinates for peak activations reported by Greene et al. were transformed to MNI space (Lacadie et al. 2008) to enable the search in Neurosynth (date of the search: 14.03.2018). Full results can be found in the following links: BA 46 (http://neurosynth.org/locations/46_36_24_6/), left BA 7/40 (http://neurosynth.org/locations/-48_-68_26_6/) right BA 7/40 (http://neurosynth.org/locations/50_-60_18_6/)

such as the ones that Greene and colleagues associate with emotion. Decades of fMRI research have provided a structure-function mapping that is far from selective, with most brain areas involved in many different mental processes (Poldrack 2010; Anderson et al. 2013). When differential activity is found in a specific brain area, it is usually not possible to determine which of the many mental processes that are (to a similar extent) likely to be engaged is actually engaged. Thus, although the probabilistic approach makes reverse inference viable in principle, in practice it is of limited use in most cases. The same seems to be true for extant comparative approaches to reverse inference, which also need strong structure-function links in order to avoid the 'compatibility problem'.

3. Lindquist et al. (2012): Psychology Project and Ontology Testing

3.1. Research context

Lindquist et al.'s 2012 study is an example of 'science by synthesis'. Instead of conducting a new fMRI experiment to investigate emotion, Lindquist and colleagues conducted a meta-analysis of the neuroimaging literature on emotion. A meta-analysis is a statistical technique that allows to obtain a formal synthesis of the results across different studies. In neuroimaging research, the difficulty of establishing selective structure-function links is one of the main motivations for conducting these formal syntheses of the existing evidence (Yarkoni et al. 2010). Meta-analyses of neuroimaging studies can serve to evaluate both the consistency and the selectivity of associations between localised brain activity and mental function. First, by aggregating data across experiments investigating the neural correlates of the *same* mental process, meta-analyses can provide information about the brain areas that are *consistently* associated with that mental process. Second, when used to aggregate data across studies investigating *different* mental processes, meta-analyses can also allow to evaluate the *selectivity* of brain areas for those mental processes.

The lack of selectivity of structure-function associations has also prompted a debate about whether we should rethink our cognitive ontology, our theory about the organization of the mind (Price and Friston 2005; Poldrack et al. 2009; Lenartowicz et al. 2010). Some authors have suggested that mental processes might not map well onto brain structures because the cognitive ontology that has guided the design of neuroimaging experiments

is not adequate. That is, the task components isolated in forward inferences might not correspond to basic operations of the mind. Neuroimaging research assumes that psychology is *not* independent from neuroscience, so it is to be expected that basic operations of the mind are also basic operations of the brain. Thus, when tasks that are supposed to isolate different mental processes produce overlapping patterns of brain activity, the nature of those processes as basic operations or building blocks of our mind is called into question.

Following this 'same pattern of brain activity equals same mental processes' logic, neuroimaging results can be used to test the adequacy of our cognitive ontology, and possibly guide a reformulation of the categories we use to understand the organization of the mind. Meta-analyses are specially well suited for this ontology testing approach, as they can tell us whether tasks that are supposed to involve similar (distinct) mental processes, recruit activity in the same (distinct) brain structures. This approach is used by Lindquist and colleagues to inform the debate around the nature of emotions. In particular, they investigate the question of whether emotions are "basic" or "constructed".

3.2. Lindquist et al.'s study

Lindquist and colleagues conducted a meta-analysis of the results from almost a hundred neuroimaging studies. They selected studies investigating the neural correlates (forward inference) of the perception and experience of five discrete emotions: anger, sadness, fear, disgust, and happiness. These have been considered to be 'basic emotions'. According to this basic emotion theory (Ekman 1999), these emotions are irreducible building blocks of the mind, which are the result of evolutionary pressures, and have their roots in hard-wired mechanisms in the brain and the body. On this view, it is to be expected for each basic emotion to be associated with activity in a specific region or network of regions in the brain[8]. That is, that there are brain regions *selectively* associated with each basic emotion (but see Scarantino and Griffiths 2011). In contrast to basic emotion theory, a constructionist view about emotion posits that all emotions emerge from the combination of the same domain-general mental processes: core affect and conceptualization (Lindquist and Barrett 2008). Thus, constructivism predicts that the patterns of brain

[8] It is necessary to posit emotion-specific central nervous system (CNS) activity in my account of basic emotions. […] There must be unique physiological patterns for each emotion, and these CNS patterns should be specific to these emotions not found in other mental activity. (Ekman 1999, page 50)

activity associated with each emotion will overlap. In other words, that there are no selective associations between brain regions and each basic emotion.

The results of Lindquist et al.'s meta-analysis shows that it is not possible to distinguish each basic emotion by its associated pattern of brain activity. Each emotion was associated with activity across a number of different brain regions, and none of those regions were selective for a particular emotion. For example, the amygdala, which has been taken to be the 'fear area', or at least the most important hub in a fear circuit, is shown to be consistently recruited by the experience and perception of all the emotions in the analysis. Similar results were found for other areas such as the anterior insula and the lateral orbitofrontal cortex, among others.

3.3. Discussion

Lindquist and colleagues' results show that there are no 'basic emotion circuits' in the brain. Instead, different basic emotions share (to some extent) the same neural substrates. Contrary to what basic emotion theory posits, this suggests that emotions are not basic operations of the mind. Furthermore, the brain regions consistently activated by emotions have also been associated with non-emotional processes. This seems to support the constructivist view, in which emotions are built from the combination of operations that are not specific to emotion, but rather domain-general. These conclusions, however, do not remain unchallenged. It is possible to question both the negative claim against basic emotions (A. Scarantino 2012), and the positive claim in favour of constructivism (Sander 2012; Scherer 2012). Regarding the second objection, it has been argued that the results of Lindquist et al.'s meta-analysis do not provide clear support for constructivism. The evidence is merely consistent with this theory, and thus also compatible with other theories of emotion such as appraisal theory. However, the evidence is *inconsistent* with basic emotion theories. Thus, the negative claim about the status of emotion categories as basic operations of the mind still holds.

Even proponents of basic emotions theory have accepted the evidence regarding the absence of basic emotion circuits in the brain (Scarantino 2012; Adolphs 2016). Similar to Lindquist and colleagues, they have used the data to argue in favour of a reformulation of our cognitive ontology, although not a constructivist one.

Scarantino (2012) agrees that there are no specific biological markers for each emotion category. However, he argues that this does not prove basic emotion theory wrong. He claims that basic emotions exist, but they do not match our traditional folk-psychological categories of emotion. The reason that there are no selective associations between brain activity and emotion categories is that neuroimaging studies of emotion have been guided by the wrong cognitive ontology. Like Lindquist and colleagues, Scarantino argues that the evidence should make us rethink the organisation of the mind. However, instead of considering that emotions are not part of an adequate cognitive ontology, he argues that *our current emotion categories* are not part of an adequate cognitive ontology.

Adolphs (2016) has argued that the main problem with the cognitive neuroscience of emotion to date has been the lack of conceptual clarity. On his view, in order to find selective associations between emotions and brain circuits, neuroimaging studies need to employ more rigorous designs. In particular, it is important to distinguish between emotion states and the conscious experience, attribution, conceptual knowledge, and expression of emotions. Only emotion states would constitute building blocks of the mind, while the others are abilities derivative from them. Neuroimaging studies usually conflate these different dimensions, and do not properly isolate or control for them. This is why neuroimaging studies have not found specific neural correlates for emotions, and not because basic emotion theory is wrong. On Adolphs' view, in order for the neuroscientific study of emotion to progress, researchers should take into account these conceptual distinctions when designing neuroimaging experiments.

In both examples above, the fMRI evidence is taken to be relevant to assess hypotheses about our cognitive ontology, and the categories we use to capture the organization of the mind are reformulated in the light of this evidence. Further meta-analysis of neuroimaging data regarding constructs other than emotion, such as working memory (Lenartowicz et al. 2010), have been used for the same ontology-testing purpose.

4. General discussion

The two case studies presented in this chapter exemplify different ways in which fMRI evidence can be used to inform philosophical work, and the problems associated with them.

The discussion of Greene et al.'s (2001) results brought out the problems associated with using fMRI to identify the mental processes involved in an experimental task. This strategy, known as reverse inference, faces a problem because associations between brain structure and mental processes are always one-to-many. There are two main proposals to make reverse inference a viable strategy: (1) to reformulate it in probabilistic terms and (2) to use it to decide between competing hypotheses. Greene et al.'s example showed that (1) usually provides little support for cognitive hypotheses and (2) is prone to fail because cognitive hypotheses are often compatible with multiple different neuroimaging results (see Russell A. Poldrack 2006; and G. Del Pinal and Nathan 2017 for similar conclusions).

Lindquist et al. (2012) provided an example of how the many-to-many character of structure-function links, which limited the use of reverse inference, can be used to inform questions about cognitive architecture. Their approach consists in looking at the degree of overlap between patterns of brain activity across tasks, to assess whether those tasks involve the same or distinct mental processes. Lindquist et al.'s example is especially powerful because their results are used against a theory about the organization of the mind that *makes predictions* about brain activity. However, it has been argued that this strategy can inform debates about the organization of the mind even when the theories involved make no predictions about brain activity (Mather and Kanwisher 2013). Although this approach implies assumptions which are debatable, such as the correspondence between brain and mental function, and the relevance of neuroscientific data for psychological science, it has greater potential than reverse inference.

4.1. Network-based approaches

A fundamental difference between Greene et al.'s and Lindquist et al.'s approaches is that the latter does not rely on localising mental processes in particular brain structures. Lindquist et al.'s conclusions against basic emotions theory depend just on the degree of overlap between emotions' neural correlates, in whichever part of the brain this overlap happens. The 'localizationist paradigm' in cognitive neuroscience has been heavily criticised. In particular, it has been argued that the subtractive method (see section 2.1) relies on certain theoretical assumptions about the modular and serial organization of the mind and the brain which are likely to be false (Orden and Paap 1997; Uttal 2001). In consequence, many researchers in cognitive neuroscience have switched from location-

based to network-based approaches. On the network-oriented view, mental function does not depend on the activity of isolated brain modules, but on complex patterns of interaction between anatomically separated neural populations. This switch of perspective requires refining our methods, to be better able to detect these kinds of interactions.

A method of increasing popularity is multivariate decoding (Hebart and Baker 2017). Multivariate decoding differs from the subtractive method in two substantial ways. First, while subtractive designs use *univariate* methods of analysis, which consist in running a separate analysis on each voxel (aggregation of neurons, see section 1.1), *multivariate* methods consist on the joint analysis of multiple voxels. This allows us to analyse distributed patterns of activity across separate neural populations. Second, while subtractive designs use methods of *encoding*, which aim to predict the neural data from the experimental task, *decoding* aims to predict the experimental task from the neural data. Multivariate decoding strategies use tools from machine learning to create classifiers which, after being 'trained' (or 'fed') with fMRI data for different tasks, can predict (decode) which task is being performed, just by looking at the associated pattern of brain activity. Multivariate decoding has been proposed as a methodological improvement on both strategies discussed here: reverse inference and ontology testing.

Some authors have claimed that the use of multivariate decoding could provide a viable alternative to traditional location-based reverse inference (Poldrack 2011; Del Pinal and Nathan 2017). For example, if we want to know whether a task T involves mental process M or mental process M', we can design a task that uncontroversially engages M, a second task that uncontroversially engages M', and train a classifier with fMRI data for both tasks. Then, we collect data for task T, and use the classifier to determine whether T involves M or M' based on the decoding accuracy. That is, based on whether the pattern of brain activity in T resembles the one in the task that involves M or the one in the task that involves M'. This pattern-based reverse inference is an interesting possibility. However, at least to date, it does not seem suitable for use within experimental philosophy's intuitions project. Multivariate decoding is especially informative with simple tasks, in which we are certain about the mental processes involved. For example, it is possible to decode whether a subject is viewing a shoe or a bottle by looking at the pattern of brain activity associated with each task (Norman et al. 2006). However, the possibilities of multivariate decoding as a base for reverse inferences are limited when using complex tasks, which are likely to involve several different processes (not just M

or M', but also M_1, M_2, M_n). In these cases, it is difficult to determine whether the decoding accuracy is due to the tasks being similar in terms of the mental process of interest (M or M'), or similar in terms of the other mental processes involved (M_1, M_2, M_n).

Multivariate decoding has also been proposed as an improvement on Lindquist et al.'s methodology and has led to doubts about their conclusions. Although Lindquist and colleagues' claim against basic emotions theory does not rely on the localisation of mental states, their meta-analysis uses data coming from location-based subtractive designs. If the neural basis for each basic emotion is to be found in distributed patterns of brain activity, then their study is fundamentally incapable of finding distinct neural correlates for each basic emotion (Hamann 2012). Using multivariate decoding strategies, other studies have shown that it is possible to discriminate between basic emotions based on their associated patterns of brain activity (see Kragel and LaBar 2016 for a review). However, it is important to note that the success of multivariate decoding has to do with predictive power, and predictive power does not necessarily imply neurobiological reality (Poldrack 2011; Hebart and Baker 2017; Ritchie et al. 2017). That a distributed pattern of brain activity can be used to predict, for example, the engagement of an emotion (*decode* the emotion), doesn't mean that that pattern of activity has the function of generating the emotion (*encode* the emotion). That pattern is not necessarily present in any of the individual emotion instances, so it cannot be considered as an emotion circuit in the brain (Clark-Polner et al. 2016). In fact, the pattern of distributed neural activity associated with each emotion differs across studies (Kragel and LaBar 2016). This is not an argument against the success of multivariate decoding in distinguishing between different emotions. But it is an argument against its significance. It suggests that successfully decoding an emotion is not the same as finding a neural circuit for that emotion. Thus, successful decoding of emotions will not provide evidence in favour of basic emotion theory.

5. Conclusion

In this chapter, I discussed the possibilities of fMRI for the purposes of experimental philosophy. In the introduction, I briefly described fMRI and the two main projects within

experimental philosophy: the intuitions project, and the psychology project. This allowed us to identify two specific ways in which fMRI can be used in experimental philosophy: reverse inference within the intuitions project, and ontology testing within the psychology project. I used two examples to discuss the methods associated with each of these approaches.

The pluripotency of brain areas, that is, the one-to-many character of associations between brain structure and mental function, was shown to limit the possibilities of reverse inference. Although methodological advances might overcome this problem, the ones available to date do not seem suitable for the purposes of experimental philosophy's intuitions project. Similar to Greene et al.'s hypothesis regarding deontological intuitions, other researchers in experimental philosophy have advanced explanations in terms of emotional biases for compatibilist intuitions (Nichols and Knobe 2007) or intuitions about the intentionality of bringing about negative side-effects (Nadelhoffer 2006). The latter hypothesis has already been tested with fMRI (Ngo et al. 2015). However, as in the case of Greene et al.'s study, their neuroimaging results are open to alternative explanations (Díaz, Viciana and Gomila 2017). One of the main conclusions of this chapter is that experimental philosophers should not use fMRI to test these kinds of hypotheses.

Ontology testing, on the other hand, was shown to have the potential to address philosophically relevant issues about the organization of the mind. Although this approach itself raises a series of philosophical questions about the relationship between neuroscience and psychology, this should be an *additional* reason for philosophers to engage in the debate. Building a proper cognitive ontology is a project that calls for the interdisciplinary approach characteristic of experimental philosophy's Psychology Project. Furthermore, questions about the nature of mental states such as pain are at the centre of much philosophical debate. Here, I suggest that fMRI evidence can be used to inform these debates. For example, it could be possible to use fMRI to investigate whether pain should be understood as a bodily sensation (like a tickle) or a complex emotional state (like disgust) by looking at the degree of overlap between the patterns of brain activity recruited by each of these processes.

To sum up, I argue that (1) fMRI is currently not suitable for the intuitions project's goal of discovering the mental processes underlying intuitions, but (2) experimental

philosophers should explore the potential of fMRI to inform questions about cognitive architecture within the psychology project.

Acknowledgments

I would like to thank Jonas Blatter, Eugen Fischer, Guillermo del Pinal, Lena Kästner, Colin Klein, Kevin Reuter and an anonymous reviewer for their comments on early drafts of this chapter.

Suggested reading

Hanson, S. J. and Bunzl, M. (2010) *Foundational issues in human brain mapping*. MIT Press.

Hebart, M. N. and Baker, C. I. (2017) 'Deconstructing multivariate decoding for the study of brain function', *NeuroImage*. Elsevier Ltd, (April), pp. 1–15. doi: 10.1016/j.neuroimage.2017.08.005.

Klein, C. (2010) 'Philosophical Issues in Neuroimaging.', *Philosophy Compass*, 5(2), pp. 186–198. doi: 10.1111/j.1747-9991.2009.00275.x.

Logothetis, N. K. (2008) 'What we can and what we cannot do with fMRI', *Nature*, 453(7197), pp. 869–78. doi: 10.1038/nature06976.

Poldrack, R. A. and Yarkoni, T. (2016) 'From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure', *Annual Review of Psychology*, 67(1), p. annurev-psych-122414-033729. doi: 10.1146/annurev-psych-122414-033729.

References

Adolphs, R. (2016) 'How should neuroscience study emotions? by distinguishing emotion states, concepts, and experiences', *Social Cognitive and Affective Neuroscience*. Oxford University Press, 12(1), p. nsw153. doi: 10.1093/scan/nsw153.

Anderson, M. L., Kinnison, J. and Pessoa, L. (2013) 'Describing functional diversity of brain regions and brain networks.', *NeuroImage*, 73, pp. 50–8. doi: 10.1016/j.neuroimage.2013.01.071.

Clark-Polner, E., Johnson, T. D. and Barrett, L. F. (2016) 'Multivoxel Pattern Analysis Does Not Provide Evidence to Support the Existence of Basic Emotions', *Cerebral Cortex*. Oxford University Press, 27(3), p. bhw028. doi: 10.1093/cercor/bhw028.

Díaz, R., Viciana, H. and Gomila, A. (2017) 'Cold Side-Effect Effect: Affect Does Not Mediate the Influence of Moral Considerations in Intentionality Judgments', *Frontiers in Psychology*, 08(February), p. 295. doi: 10.3389/fpsyg.2017.00295.

Ekman, P. (1999) 'Basic Emotions', *Handbook of cognition and emotion*, pp. 45–60. doi: 10.1017/S0140525X0800349X.

Fischer, E. and Collins, J. (2015) 'Rationalism and naturalism in the age of experimental philosophy', in Fischer, E. and Collins, J. (eds) *Experimental Philosophy, Rationalism, and Naturalism. Rethinking Philosophical Method*. Routledge, pp. 3–33. Available at: https://books.google.es/books?hl=en&lr=&id=l1uhCAAAQBAJ&oi=fnd&pg=PT10&dq=rationalism+and+naturalism+in+the+age+of+experimental+philosophy&ots=jHJ1zwkG4s&sig=Q3AVbXBHOcIPaEwrC1RDxShIfns (Accessed: 21 July 2017).

Greene, J. D. *et al.* (2001) 'An fMRI Investigation of Emotional Engagement in Moral Judgment', *Science*, 293(5537), pp. 2105–2108. doi: 10.1126/science.1062872.

Greene, J. D. (2014) 'Beyond Point-and-Shoot Morality : Why Cognitive (Neuro)Science Matters for Ethics', *Ethics*, 124(4), pp. 695–726. doi: 10.1086/675875.

Hamann, S. (2012) 'What can neuroimaging meta-analyses really tell us about the nature of emotion?', *Behavioral and Brain Sciences*. Cambridge University Press, 35(03), pp. 150–152. doi: 10.1017/S0140525X11001701.

Hebart, M. N. and Baker, C. I. (2017) 'Deconstructing multivariate decoding for the study of brain function', *NeuroImage*. Elsevier Ltd, (April), pp. 1–15. doi: 10.1016/j.neuroimage.2017.08.005.

Huebner, B., Dwyer, S. and Hauser, M. (2009) 'The role of emotion in moral

psychology', *Trends in Cognitive Sciences*. Elsevier Ltd, 13(1), pp. 1–6. doi: 10.1016/j.tics.2008.09.006.

Klein, C. (2010) 'Philosophical Issues in Neuroimaging.', *Philosophy Compass*, 5(2), pp. 186–198. doi: 10.1111/j.1747-9991.2009.00275.x.

Klein, C. (2011) 'The Dual Track Theory of Moral Decision-Making: a Critique of the Neuroimaging Evidence', *Neuroethics*. Springer Netherlands, 4(2), pp. 143–162. doi: 10.1007/s12152-010-9077-1.

Knobe, J. and Nichols, S. (2017) 'Experimental Philosophy', *Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/archives/win2017/entries/experimental-philosophy/ (Accessed: 6 January 2018).

Kragel, P. A. and LaBar, K. S. (2016) 'Decoding the Nature of Emotion in the Brain', *Trends in Cognitive Sciences*. Elsevier Ltd, 20(6), pp. 444–455. doi: 10.1016/j.tics.2016.03.011.

Lacadie, C. M. *et al.* (2008) 'More accurate Talairach coordinates for neuroimaging using non-linear registration', *NeuroImage*, 42(2), pp. 717–725. doi: 10.1016/j.neuroimage.2008.04.240.

Lenartowicz, A. *et al.* (2010) 'Towards an ontology of cognitive control.', *Topics in cognitive science*, 2(4), pp. 678–92. doi: 10.1111/j.1756-8765.2010.01100.x.

Lindquist, K. a. and Barrett, L. F. (2008) 'Constructing emotion: The experience of fear as a conceptual act', *Psychological Science*, 19(9), pp. 898–903. doi: 10.1111/j.1467-9280.2008.02174.x.

Lindquist, K. a *et al.* (2012) 'The brain basis of emotion: a meta-analytic review.', *The Behavioral and brain sciences*, 35(3), pp. 121–143. doi: 10.1017/S0140525X11000446.

Logothetis, N. K. (2008) 'What we can and what we cannot do with fMRI', *Nature*, 453(7197), pp. 869–78. doi: 10.1038/nature06976.

Machery, E. (2013) 'In Defense of Reverse Inference', *The British Journal for the Philosophy of Science*, 65(2), pp. 251–267. doi: 10.1093/bjps/axs044.

Machery, E. (2017) *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.

Mather, M. and Kanwisher, N. (2013) 'How can fMRI inform cognitive theories', 8(1), pp. 108–113. doi: 10.1177/1745691612469037.How.

Mitterschiffthaler, M. T. *et al.* (2007) 'A functional MRI study of happy and sad affective states induced by classical music', *Human Brain Mapping*, 28(11), pp. 1150–1162. doi: 10.1002/hbm.20337.

Mole, C. and Klein, C. (2010) 'Confirmation, Refutation, and the Evidence of fMRI'. Available at: http://philpapers.org/rec/MOLCRA (Accessed: 8 July 2015).

Moll, J., Eslinger, P. J. and De Oliveira-Souza, R. (2001) 'Frontopolar and anterior temporal cortex activation in a moral judgment task: Preliminary functional MRI results in normal subjects', *Arquivos de Neuro-Psiquiatria*, 59(3 B), pp. 657–664. doi: 10.1590/S0004-282X2001000500001.

Nadelhoffer, T. (2006) 'Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality', *Philosophical Explorations*, 9(2), pp. 203–219. doi: 10.1080/13869790600641905.

Ngo, L. *et al.* (2015) 'Two Distinct Moral Mechanisms for Ascribing and Denying Intentionality', *Nature Scientific Reports*, 5, pp. 1–11. doi: 10.1038/srep17390.

Nichols, S. and Knobe, J. (2007) 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions', *Nous*. Blackwell Publishing Inc, 41(4), pp. 663–685. doi: 10.1111/j.1468-0068.2007.00666.x.

Norman, K. A. *et al.* (2006) 'Beyond mind-reading: multi-voxel pattern analysis of fMRI data', *Trends in Cognitive Sciences*, 10(9), pp. 424–430. doi: 10.1016/j.tics.2006.07.005.

Orden, G. C. Van and Paap, K. R. (1997) 'Functional Neuroimages Fail to Discover Pieces of Mind in the Parts of the Brain', *Philosophy of Science*, 64(S1), p. S85. doi: 10.1086/392589.

Parens, E. and Johnston, J. (2014) 'Neuroimaging: beginning to appreciate its complexities.', *The Hastings Center report*, Spec No, pp. S2-7. doi: 10.1002/hast.293.

Del Pinal, G. and Nathan, M. J. (2013) 'There and up again: on the uses and misuses of neuroimaging in psychology.', *Cognitive neuropsychology*, 30(4), pp. 233–52. doi: 10.1080/02643294.2013.846254.

Del Pinal, G. and Nathan, M. J. (2017) 'Two Kinds of Reverse Inference in Cognitive Neuroscience', *The Human Sciences after the Decade of the Brain*, pp. 121–139. doi: 10.1016/B978-0-12-804205-2.00008-2.

Poldrack, R. A. (2006) 'Can cognitive processes be inferred from neuroimaging data?', *Trends in Cognitive Sciences*, 10(2), pp. 59–63. doi: 10.1016/j.tics.2005.12.004.

Poldrack, R. A. (2010) 'Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed?', *Perspectives on psychological science : a journal of the Association for Psychological Science*, 5(6), pp. 753–761. doi: 10.1177/1745691610388777.

Poldrack, R. A. (2011) 'Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding', *Neuron*. Elsevier Inc., 72(5), pp. 692–697. doi: 10.1016/j.neuron.2011.11.001.

Poldrack, R. A., Halchenko, Y. O. and Hanson, S. J. (2009) 'Decoding the large-scale structure of brain function by classifying mental states across individuals', *Psychological Science*, 20(11), pp. 1364–1372. doi: 10.1111/j.1467-9280.2009.02460.x.

Price, C. J. and Friston, K. J. (2005) 'Functional ontologies for cognition: The systematic definition of structure and function', *Cognitive Neuropsychology*, 22(3–4), pp. 262–275. doi: 10.1080/02643290442000095.

Ritchie, J. B., Kaplan, D. and Klein, C. (2017) 'Decoding The Brain: Neural Representation And The Limits Of Multivariate Pattern Analysis In Cognitive Neuroscience', *bioRxiv*, p. 127233. doi: 10.1101/127233.

Rose, D. and Danks, D. (2013) 'In Defense of a Broad Conception of Experimental Philosophy', *Metaphilosophy*, 44(4), pp. 512–532. doi: 10.1111/meta.12045.

Roskies, A. L. (2008) 'Neuroimaging and Inferential Distance', *Neuroethics*, 1(1), pp. 19–30. doi: 10.1007/s12152-007-9003-3.

Sander, D. (2012) 'The role of the amygdala in the appraising brain', *Behavioral and Brain Sciences*. Cambridge University Press, 35(03), p. 161. doi: 10.1017/S0140525X11001592.

Scarantino, a. (2012) 'How to Define Emotions Scientifically', *Emotion Review*, 4(4), pp. 358–368. doi: 10.1177/1754073912445810.

Scarantino, A. (2012) 'Functional specialization does not require a one-to-one mapping between brain regions and emotions', *Behavioral and Brain Sciences*. Cambridge University Press, 35(03), pp. 161–162. doi: 10.1017/S0140525X11001749.

Scarantino, A. and Griffiths, P. (2011) 'Don't give up on basic emotions', *Emotion Review*, 3(4), pp. 444–454. doi: 10.1177/1754073911410745.

Scherer, K. R. (2012) 'Neuroscience findings are consistent with appraisal theories of emotion; but does the brain "respect" constructionism?', *Behavioral and Brain Sciences*. Cambridge University Press, 35(03), pp. 163–164. doi: 10.1017/S0140525X11001750.

Uttal, W. R. (2001) 'The New Phrenology:The Limits of Localising Cognitive Processes in the Brain', pp. 221–228.

Yarkoni, T. *et al.* (2010) 'Cognitive neuroscience 2.0: Building a cumulative science of human brain function', *Trends in Cognitive Sciences*, 14(11), pp. 489–496. doi: 10.1016/j.tics.2010.08.004.

Yarkoni, T., Poldrack, R. A. and Nichols, T. (2011) 'Large-scale automated synthesis of human functional neuroimaging data', *Nature Methods*, 8(8), pp. 665–670. doi: 10.1038/nmeth.1635.Large-scale.