From 'blobs' to mental states

Author: Javier Gomez-Lavin

DOI: 10.5040/9781350349513
Page Range: 77–102

The epistemic successes and limitations of functional magnetic resonance imaging

Javier Gomez-Lavin

Table of Contents

1 From brains to 'blobs': The basics of functional magnetic resonance imaging (fMRI)

How do we go from placing people in the bore of a 7,000-kg superconducting magnet to generating pictures of their brain laced with 'colorful hot spots of activity' (Roskies, 2007, p. 860)? That we can manage this feat of opening a non-invasive window unto the brain, and indeed the fact that most researchers can train students to run experiments using these tools in mere months, is awe-inspiring. Although approaching any level of mastery with this technique can take years of training,[1] we need to only review some of the basics of nuclear magnetic resonance, image encoding and the blood-oxygen-level-dependent (or 'BOLD') response to appreciate the epistemic opportunities – and challenges – that this tool poses for the philosophically inclined.

The forerunner to fMRI is nuclear magnetic resonance imaging (NMR) whose development can be traced through the second half of the twentieth century.[2] Much as a viola can be thought of *resonating* as its vibrating strings produce soundwaves which are greatly amplified by the body of the instrument to generate at times even a loud musical note, when groups of atoms begin to 'vibrate' together they too can change their properties in a way that resembles this amplification process, although they feature changes in their net electromagnetic properties as opposed to changes in pressure that we experience as sound (R. Poldrack, 2018). How do we get these atoms to 'sing'? Simply put, we expose them to an enormous magnetic field, with most scanners subjecting tissues or individuals to magnetic fields five orders of magnitude greater than what the Earth manages on average.[3] Subjecting atoms, specifically their protons (and to be clear, here we're largely concerned with the hydrogen atoms which make up the majority of the atoms in the water molecules that themselves make up a good majority of the molecules within your cells and tissues), to such strong magnetic fields aligns them to the local field produced by the scanner (Uttal, 2001, p. 76). Protons, as with most elementary particles, have a further property which physicists have analogized to and term 'spin', and for our purposes we can think of a field of protons as each spinning along the strong magnetic field with which they're aligned (Buxton, 2012, p. 5). However, these protons wobble around their axis of rotation a bit like how a child's spinning top processes as it loses speed and begins to topple. It's this collective precession of the spinning protons that 'acts a radiating antenna' and emits low-frequency radio waves that we can detect through a number of strategically placed coils of wire, much in the same way that one would pick up an analogue radio or TV signal (Uttal, 2001, p. 77). It's in this sense that we're using *nuclear resonance*, as we're causing the constituents of the nucleus of atoms (i.e. protons) to resonate.

At this point all we have is a person stuck in the bore of a colossal superconducting magnet with their hydrogen atoms tuned by the magnet into emitting synchronized radio waves along with a negligible amount of heat. We still need a few further components to get from this point to an image of the brain in action. How do we produce *images* in the first place? For that we need *contrast* (Ogawa, 2012, p. 608). That is, we need to exploit some inconsistency inherent in the material that we're interested in, where we know that that inconsistency correlates with the structures or dynamic processes we're trying to image. To do this, we generate a second (or sometimes several additional) pulse of radio waves and direct it at the synchronized protons. This pulse briefly knocks the protons out of their previous alignment with the large local magnetic field produced by the scanner and in doing

so alters their 'wobble' (Roskies, 2007, p. 863). In turn, these changes, and the *rates* that protons take to return to the alignment and spin induced by the large local field, affect the resultant radio frequency signal. In effect, it's these further pulses that transform the subject in the scanner from a human radio antenna blasting a single note into something resembling a tune.

Luckily for us, and much to the chagrin of the radiologists who must master the variability on offer, tissues differ in how their protons return to alignment with the large local field. Specifically, protons feature a number of 'time constants' that capture this rate along two dimensions with 'T1', or the longitudinal time constant representing the magnetic changes protons undergo as they realign in the direction of the local field, and 'T2', the transverse time constant which captures the realignment in a plane perpendicular to the magnetic field (Uttal, 2001, p. 79). The physical basis and specifics of T1 and T2 can quickly become overwhelming (consult, for instance, Buxton, 2012); however, for our purposes all that is relevant is that we can exploit the variability in T1 and T2 (themselves caused by a complex constellation of factors including water distribution and density) rates to generate an anatomical image of the specimen in question.

As tissues can have different T1 and T2 values, so too can dynamic processes in the body have distinct time constant values. Thanks to, again, a complex of physical and chemical factors, it turns out that as the haemoglobin molecule responsible for carrying oxygen to your tissues loses its four O_2 molecules it becomes *paramagnetic,* or weakly magnetic (Buxton, 2012, p. 2). Although the magnetic difference between haemoglobin and *deoxy*haemoglobin had been known earlier in the twentieth century, it was only in the 1990s that Ogawa and others began to use this property in magnetic resonance imaging (Bandettini, 2012; Ogawa, 2012). Changes in the local magnetic field brought about by the relative mix of oxygenated to deoxygenated blood result in a local inhomogeneity that can be detected as protons realign to the large magnetic field induced by the scanner, and its time constant, termed 'T2*', that allows us to estimate local changes in the amount of oxygenated blood present (Buxton, 2012, p. 2; Uttal, 2001, p. 85). It's through the T2* time constant, along with the adoption of more specialized imaging techniques – including so-called, 'single-shot' echo planar imaging – that we're able to quantify the *BOLD* response central to the *functional* aspect of fMRI (Kwong, 2012).

By now it should be relatively clear how we can use a giant magnet to induce changes in endogenous radio frequencies that allow fMRI, almost like a sonar device, to peer inside your body and create images of your respective organs and even dynamic processes that are ongoing within those organs, such as the relative fraction of oxygenated blood present at a given time. However, you might well wonder just how we're able to piece together *where* a given radio frequency signal from your body is coming from. After all, if the scanner induces your protons to resonate, and if a typical body has something in the order of 10^{27} (give or take) water molecules, then how on Earth do we begin the process of determining which molecules in space are the prime contributors to a given radio signature? Roughly, as Russ Poldrack puts it, the scanners we currently have are capable of subtly varying the field they induce along the length of the specimen, and by combining this subtle variation with radio frequency pulses that are identified or 'encoded' with information we're able to reconstruct a relatively accurate three-dimensional map of where endogenously produced radio waves are coming from (R. Poldrack, 2018). When speaking of *f*MRI, we term these three-dimensional volumes that constitute the image, 'voxels', which are approximately 3-mm cubes of space (although finer-grained spatial estimates can be achieved with larger magnetic fields and other techniques; cf. Goebel, 2012).

It's by overlaying the 'functional images' we generate from the T2* time constant, namely the BOLD response, onto an anatomical image, often constructed from a separate slower T1-focused scan of an individual's brain, that we can paint those 'colorful spots of activity' that are the hallmarks of fMRI and which can be seen in Figure 4.2 in Section 4 (this process is largely mediated via specialized software applications such as BrainVoyager, cf. Goebel, 2012; Roskies, 2007, p. 860).

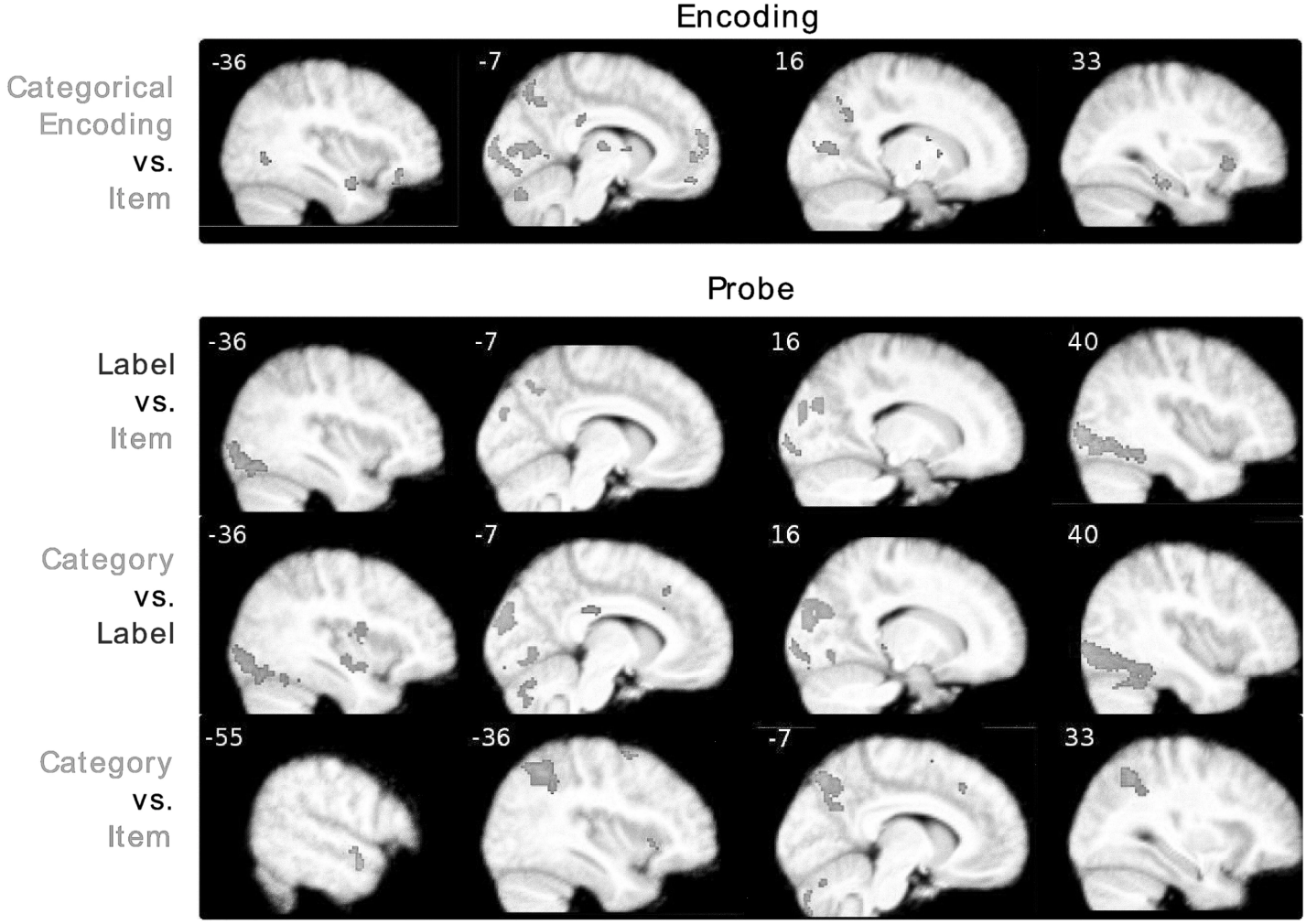

Figure 4.2 'Whole-brain univariate analyses: activity differing between conditions within individual trial epochs (Encoding and Probe). Top figure: Encoding epoch. Red: Categorical Encoding (Category and Label trials) greater than Item; blue: Item greater than Categorization Encoding. Bottom figures: Probe epoch. Top: Green: Label greater than Item; blue: Item greater than Label. Middle: Green: Label greater than Category; Red: Category greater than Label. Bottom: Red: Category greater than Item. Blue: Item greater than Category. Regions of activity are overlaid on the average normalized anatomical image across subjects. For each contrast, we generated maps at an uncorrected threshold of $p < 0.001$ and corrected for multiple comparisons using the topological false-discovery rate' (Braunlich et al., 2015, p. 150). Reproduced from *NeuroImage,* vol. 107, Braunlich et al., 'Frontoparietal networks involved in categorization and item working memory'. © 2015, with permission from Elsevier

By now you might start to recognize some of the most basal limits and benefits of fMRI. Thanks to the mapping procedure described earlier, it has an excellent spatial resolution. Assuming your subject isn't moving too much while they attempt to lie completely still in the bore of a deafeningly loud and claustrophobia-inducing superconducting magnet, and assuming you've properly overlaid the images (which often requires the averaging and smoothing of many subjects' anatomical data into a single 'Talairach' space which serves as a kind of standardized atlas of the human brain), then you can be reasonably sure that a given BOLD signal comes from a specific 3 mm^3 chunk of brain (R. Poldrack, 2018). However, blood, as anyone who's gotten a small nick or paper cut on a cold day can attest, takes a few seconds to start flowing. The BOLD response is thus a *delayed* response, as oxygenated blood can take a few, usually in the order of three to five, seconds to begin to reach the area in highest demand to give up its precious cargo (Buxton, 2012; Poldrack, 2018; Roskies, 2007; Singh, 2012). Hence, fMRI has poor temporal resolution of neural activity, especially when contrasted with other techniques that measure primarily electrical changes in the cortex, such as electroencephalography (EEG).

But how are we so sure that what we're imaging is neural *activity*? After all, when we generate a 'functional image' what we're really representing through our choice of colourful 'blobs' are *statistical* properties associated with a given region of space, namely those 3 mm^3 voxels of neural tissue (Klein, 2010). The hope is that those statistical properties, usually generated by interpreting the BOLD signal via a model of the *hemodynamic response function* that is indexed to the type of task that participants are undergoing, genuinely track some change in the brain that is substantially correlated with neural effort. To answer this question, we can follow the chain of reasoning elegantly laid out in Uttal's 2001 book on the subject, which I quote in full here:

> To summarize, like the PET scan, fMRI provides a means of indirectly estimating the amount of neural activity in regions of the brain, working from a logical chain of correlations. Additional brain activity leads to higher glucose metabolism, which leads to high oxygen demand, which leads to higher blood flow. As the oxygen in the newly arrived blood is used up, the hemoglobin changes from a form that is not magnetically susceptible to one that is. The change in T2* values as oxygenated hemoglobin becomes deoxygenated can be measured and the data converted to an image showing the functional, as opposed to the anatomical, properties of the human brain. (Uttal, 2001, p. 86)

Intuitively, this process makes sense: brain work is hard work and it takes a substantial amount of energy which is provided by the process of cellular respiration requiring the interaction of glucose and oxygen, hence, with more work we require more oxygen.[4] If we have a measure of the rate of oxygen use, then we can correlate it to the amount of neural effort, or activity, that a given voxel is experiencing.

Say we grant that there is some connection between neural effort and oxygen consumption, which seems reasonable enough. Are we able to justify the further claim that the BOLD response is an informative or adequate proxy for neural activity? Roskies (2007) notes that the BOLD signal may 'reflect a variety of different changes in the brain' including 'subthreshold activity, simultaneous excitation and inhibition . . . modulatory inputs from other areas . . . [and] changes in neural synchrony' (p. 866). Further, Klein (2010) argues that the images produced by fMRI 'are not maps of activation per se', but instead they provide statistically vetted clues about where in the brain we might find more evidence for a given functional hypothesis (p. 275). Putting possibly the most pessimistic spin on things, Singh (2012) argues that 'the very phrase neural activity is itself a rather poorly specified and ultimately meaningless term . . . within the cortex there are multiple neural signals . . . that might all contribute to the metabolic demand that then drives the BOLD signal' (p. 1122). We will revisit these points in the following sections, particularly as they relate to the risks of *type one* (a false positive result) and *type two* (a false negative) errors in fMRI (Poldrack, 2018, p. 61). Ultimately, the question of whether and to what extent fMRI images actually show or are tied to neural *activity* will shadow the rest of this chapter and has fuelled much of the philosophical commentary on fMRI over the past twenty plus years (e.g. Coltheart, 2006; Fodor, 1999; Glymour & Hanson, 2016; Hardcastle & Stewart, 2002; Klein, 2010; McCaffrey & Danks, 2022; Poldrack, 2010; Roskies, 2007). Effectively tackling it will first require that we learn a bit more about the kinds of tasks and paradigms that are often used in fMRI studies.

2 So you want to run an fMRI study: A historical review of methods and analyses

A rule of thumb in the sciences of the mind is that new tools make their debut paired with old methods, stimuli and tasks. Similarly, we can trace the conceptual roots of the earliest batch of fMRI studies from their then recent PET (or positron emission tomography, a separate branch of metabolic neuroimaging) predecessors all the way back to at least the nineteenth century with Donders's 'method of cognitive subtraction' alongside the assumption of 'pure insertion' (Courtney, 2012, p. 1186; McCaffrey & Danks, 2022). The method of subtraction, as its name implies, holds that we can deduce the time required to perform a given task, *x*, by subtracting a series of tasks featuring it from a series with additional tasks, say *x* and then *y* and *z*. In turn, we can make some inferences about the relative amount of (cognitive) effort involved in a given task. Supporting this method is the assumption of 'pure insertion', or the rather rigid view that each task is treated individually (and processed in a subsequently serial fashion) and does not affect other tasks adjacent to it in time. As a bit of a caricature, imagine that you have to prepare guacamole: you could choose to chop all the ingredients prior to mashing and combining them together or you could interleave the tasks (which would produce better results, trust me).

These assumptions hailing from the halcyon days of experimental psychology helped structure the characteristic 'blocked' design of early fMRI experiments. These experiments, which were often lifted directly from earlier PET set-ups, would require subjects to perform the same initial task, over and over, before being asked to perform a second task, again repeating the task many times (Clark, 2012; Courtney, 2012; Huettel, 2012; Uttal, 2001). For instance, subjects would be asked to tap their finger for thirty seconds at a time prior to resting for thirty seconds (Huettel, 2012, p. 1152). It's sensible to use a blocked design for many PET experiments, as PET measures the time course of glucose (or oxygen) consumption via the radioactive emissions of an injected tracer; as such, it offers good metabolic resolution (especially when, for instance, attempting to ascertain the location of an energy- hungry tumour) at the cost of temporal resolution. However, the use of blocked designs in fMRI largely limits the kinds of cognitive processes that can be studied as individuals will quickly come to expect the next stimulus within a block, and boredom or perseveration can ensue (Clark, 2012, p. 1192).

In the late 1990s 'event-related' designs began to gain ground as newer techniques for interpreting the BOLD response became more common and epistemic questions about the nature of the BOLD response, particularly how it responds to additional stimuli, were clarified (Huettel, 2012, p. 1152). Specifically, the BOLD response seems proportional to the strength of the neural activity

and shows 'superposition', wherein the BOLD response for a longer stimulus could be modelled by summing the responses measured to a series of shorter stimuli (Huettel, 2012, p. 1154). Given the linear properties of the BOLD response it also became important to vary the time between certain events and their 'epochs', or the series of events that delineate methodically relevant portions of a task trial, and this variation – often between discreet trials – is colloquially referred to as 'jittering' (Courtney, 2012, p. 1188). These designs allow experimenters to string a number of tasks together, provided that they have accurate information about the time course during which cues, stimuli, tasks and probes are presented to a given subject. As such, a much broader array of paradigms, including typical working memory tasks such as the n-back or delayed match-to-sample tasks which we'll return to in the following sections, could be studied using fMRI. Often, in event-related designs, the time course of a given method is convolved with a model of the BOLD response, and this information can then be treated as a predictor in a multiple regression model (Courtney, 2012). In turn, these event-related designs and associated univariate analyses helped to drive the expansion of fMRI at the start of the 2000s.

Another important contributor to this growth was the development of the *resting state* experiment (Poldrack, 2018; Snyder & Raichle, 2012). As their title suggests, these studies examine endogenously generated neural activity from individuals who are in some intuitive sense 'at rest', or not performing an occurrent experimental task while in the scanner. While it's fallacious to presume that the brain is ever *truly* at rest, at least as long as one's alive, and this point is taken up at length by Klein (2014); however, this isn't as significant a point of friction between philosophers of neuroscience and neuroscientists as it may seem, as experimentalists agree that 'the resting state is not truly a resting state at all' (Snyder & Raichle, 2012, p. 904). In fact, one can trace a commitment to the importance of the brain's self-generated activity back to at least Hans Berger, who in 1929 quipped about the then new EEG experiments that 'mental work . . . adds only a small increment to the cortical work which is going on continuously' (cited in Snyder & Raichle, 2012). However, all this focus on *rest* is a bit of a red herring – as the interesting finding at the heart of these studies has to do with what they reveal about the brain's *functional connectivity* or how activities across distinct and distributed brain networks vary with each other over time (R. Poldrack, 2018).

A key insight motivating the field of resting-state fMRI came from Biswal (2012) who, while studying the somatomotor system in the 1990s, noticed that activity associated with a voxel in the left motor cortex was substantially correlated with activity in the entire right motor cortex. While this correlated pattern of activity makes sense when subjects are asked to coordinate their movements (e.g. alternately tapping their right and then left index finger), he found that the pattern held even when the subjects were told to relax in the scanner (Poldrack, 2018). These findings suggest that wholesale distributed networks were active in the brain despite external task demands, and they opened up a host of questions about the scope and dynamical properties of these networks, of which there are at least a half dozen or more depending on one's preferred method and level of analysis (Lv et al., 2018). Probably the most well studied of these networks is the 'default mode' network, which seems to be involved in non-goal-directed mind-wandering – a bit like what happens when you're driving on a familiar route and you find that you're preoccupied thinking through those various emails left to send and exactly where you might go during your next holiday only to suddenly come to and realize that you've made it to your driveway (Raichle, 2015). It's this network that is thought to reduce its activity when you switch into a more attentive state during the course of executing an occurrent cognitive task.

It's interesting to note that intrinsic and endogenous changes in the BOLD response in the absence of external task were noticed in the 1990s but were often dismissed as imaging artefacts or 'noise' generated by external causes, such as subject movement (Snyder & Raichle, 2012, p. 905). That is, although experimenters were aware that there were spontaneous changes in the hemodynamic response in fMRI experiments, they rejected that these data were indicative of a result or genuine phenomenon. Though this mirrors the logic of a *type two error* where we mistakenly reject that some pattern of activity might instead represent a meaningful result, it's not quite fair to suggest that these earlier researchers were in fact committing such a mistake, since in a very real sense the alternative hypothesis that there could be meaningful endogenously generated changes in the BOLD response did not factor into their analyses. Rather and surprisingly, such a pattern better fits the mould of 'philosophers' syndrome', in that it trades a failure of imagination for an implicit claim about the nature of the neural response (Dennett, 1991). I highlight this here as a lesson in epistemic humility and of the value of imagination which might still yet pay dividends in our future studies of the brain. Finally, although resting-state studies might provide a more 'bottom-up' and less theory-laden programme directed at carving apart the brain's intrinsic neural networks, they may still, as McCaffrey and Danks (2022) point out, unhelpfully lump together smaller, finer-grained networks and in doing so import a new kind of statistical artefact (p. 585).

Just as we've seen an expansion in the kinds of methods that fMRI studies can host, so too can we chart a parallel development in the range of statistical techniques that can be applied to parse the BOLD response. In the earlier days of blocked designs researchers could simply subtract the BOLD response in a given region of interest (ROI) between the two conditions to create a functional map of the net change in BOLD response across the two tasks (Formisano & Kriegeskorte, 2012). Even correcting for multiple comparisons, since we are – even with a small ROI – comparing a change in the signal across many dozens or hundreds of voxels, what we're doing in effect in these cases is a series of simple t-tests across time and task condition. These kind of early univariate analyses – which assume the independence of the BOLD signal across individual voxels – were the hallmark of the

spreading influence of 'blobology' in the 1990s that saw researchers competing to localize the neural correlates of an array of cognitive processes (e.g. the fear area or the cat-sensitive area of the cortex, to caricature things a bit) (Poldrack, 2010, 2012). As Poldrack elegantly puts it, 'the goal of finding blobs in a significant region can drive researchers into analytic gymnastics in order to find a significant blob to report' (Poldrack, 2012, p. 1217). More plainly still, the temptation to find 'where' processes were localized in the brain often drove teams to use spurious methodological and statistical techniques, which we'll review in more depth in the next section.

The advent of event-related designs was shadowed by more complex univariate analyses, including the multiple regression models mentioned by Courtney (2012) which used the time course of the task epoch (e.g. when the cue, stimulus, delay period, probe, etc., were presented to the subject) as independent variable in a model of the BOLD response. These and other general linear models (GLMs), which will feature in the study we'll review later, attempt to parse apart the noisy raw BOLD signal in a given voxel by applying a series of informative regressors (e.g. the time course data) and controlling for error and covariate terms. Again, the focus on individual voxels, and zooming out a bit, on *regions* of interest, sits at the core of these univariate analyses and corresponding experimental designs. This makes sense if one is committed, implicitly or explicitly, to a strong metaphysics of neural localization that approaches a kind of neo-phrenology, where certain mental processes are realized in specific parts of the brain. Haxby (2012) terms this the 'strong modularity hypothesis' to which univariate analyses are specifically keyed to answer. If you're at all suspicious of this thesis, or if you're open to more distributed models of the brain's functional organization, then you might be tempted to exit the bus here and now. However, the development of experiments focused on the brain's functional connectivity, such as the resting-state studies we reviewed earlier, helped – in part – to spur other analyses of the hemodynamic response that were less concerned with the absolute magnitude of the response in a given voxel or ROI and, instead, looked to changes in the *patterns* of activation measured by fMRI across different tasks.

Consider for a moment the *tremendous* quantity of data that we throw out when performing a run of the mill univariate subtraction design: even constraining our case to a small ROI of a hundred voxels (e.g. the right fusiform face area), and assuming a conservative sampling rate of two seconds per sample per voxel and two trial runs totalling twenty minutes we will generate a matrix containing 60,000 data points (or something on the order of 30,000 paired comparisons). In order to even begin the process of combing through the data, we must assume a very conservative significance threshold (e.g. $p < 0.001$) or risk an overwhelming number of false positive results. And by a scrupulous application of this process we may, indeed, find that some region, such as the right fusiform face area, shows a heightened response for a given stimulus, such as faces. But in doing so we have to accept the trade-off between type one and type two errors, where we reject plausible effects because of the statistical threshold we've chosen. So, while the fusiform face area does show selective activation for faces, it may *also* demonstrate a substantial (but not significant) response for *other* stimuli, but this data finds itself often swept under the rug as there are sure to be *other* regions more selective for said stimuli. In effect, by discarding reams of less-than-significant data in univariate methods we recreate the 'publication bias' effect where only the most significant findings see the light of day, only in a miniature form.

Haxby (2012) illustrated how we might be able to sidestep these and the earlier implications inherent in 'strong modularity' views of the brain's functional organization by instead contrasting patterns of activity across larger areas of the brain (Poldrack, 2018). Though there are number of related techniques, *multivoxel pattern analysis* (MVPA) is dominant. As Haxby puts it, MVPA increases 'the amount of information that can be decoded from brain activity, in contrast to simpler univariate measure that indicate the extent to which a cortical field or system is globally engaged' (2012, p. 852). Properly using MVPA requires that neuroimagers shift their conceptual outlook from one primarily interested in questions surrounding the localization of neural function to one attuned to a more complicated, and ultimately messier, picture of the brain's dynamics and organization – one where a given piece of neural tissue might play a number of functional roles that themselves are dependent on the broader neurological context and task demands (Haxby, 2012).

An overly simplistic example of an MVPA study is to *split* trials and their data into two sets and then to determine whether a pattern cortical activation from one set elicited by a given stimulus could be used to predict the stimulus given to that individual based upon the first pattern's similarity to a new pattern from the second set (Poldrack, 2018). Putting a finer point on it, suppose that we show pictures of chairs and houses to our subjects while they lie in the scanner and then divide the data into two sets combining both stimuli; can we then take the averaged brain-wide pattern of chair-associated activity from the first set and use this pattern to successfully predict whether a subject *actually* saw a chair given the second, independent, set data? If your *multivoxel pattern* of chair-related activity obtained from an independent set of data can be used to accurately predict chair presentation in the second set, then there's some robust sense in which you've captured an important aspect about *how* chair-related stimuli are encoded – or perhaps even *represented* – in patterns of activation in the brain. It's this train of inferences that motivates claims about fMRI allowing us to 'read people's minds', since it's by using MVPA and other related analyses, now greatly assisted with machine learning techniques and classifiers trained with independent fMRI data (Takagi & Nishimoto, 2022), that we can seemingly and with great accuracy predict and even reconstruct the stimuli that individuals previously experienced solely on the basis of multivoxel patterns of activity.

However, this train of inference might also strike some as suspiciously similar to the maligned case of *reverse inference*, where we take neural activation to stand as evidence for the recruitment of a specific mental function (Poldrack, 2006). Contrast this to a *forward inference*, where we know (or strongly suspect) that a subject is performing a given task, say a working memory task, and then we measure the BOLD response in a given area and conclude that some area demonstrates greater response in the working memory task than in some other task and in doing so 'locate' a working memory selective region, say the dorsolateral prefrontal cortex. When at some later point we find an increased (and recalling our previous discussion of univariate analyses, statistically significant!) response in the dorsolateral prefrontal cortex to some new task, for instance a task where someone must inhibit a habitual action, and when we further use this response as evidence to claim that working memory must be involved in this new task, then we've committed the sin of reverse inference. After all, it could be that the same neural tissue is responsible for two wholly distinct tasks! Though some philosophers have defended a limited use of reverse inference in neuroimaging (Machery, 2014) and others have cast doubt on the very idea that fMRI measures psychological processes (Glymour & Hanson, 2016), it's safe to say that a consensus view is that reverse inference is best avoided when other inferential practices are ready and available for the interpretation of fMRI data. Returning to our question at the outset of this paragraph, then, why should MVPA get a free pass despite clearly fitting the template left by reverse inference? As Poldrack (2018) highlights, it's because with MVPA we are using a statistical model, specified a priori, to justify the inference, whereas in our toy case earlier, the inference rests more on an intuitive process borne from the mind of the neuroimager. Minds are notoriously subject to biases, at least slightly more so than preregistered statistical models.

To summarize then, we've seen how fMRI has gone from parroting the simpler designs and statistical models of its predecessors to accommodating the very bleeding edge of advances in machine learning, with a concomitant increase in the complexity of its methods and analyses. At the same time our tour has revealed a number of *epistemic pitfalls* that we should remain attuned to as we review other fMRI studies. These include *type one* and *type two* errors, or instances where we falsely accept a result or where we falsely reject a result, respectively, the background assumption of *strong modularity* and its counterpart of functional localizability that entail a constrained mapping between mental function and neural tissue and the sin of reverse inference borne from our biases that tempt us to presume the presence of a cognitive function from the selective activation of an ROI. We've already seen the ill effects of these pitfalls in action and may even see one or two of them resurface in our review of the fMRI study in a few sections; however, before we begin that process we'd do well to pause and take stock of a few other epistemic hurdles that promise to trip up an aspiring neuroimager.

3 A miscellany of further epistemic pitfalls

Thanks to our tour of the history of fMRI and its methods we've seen many a serious epistemic pitfall emerge, including the risks of *type one* and *type two* errors, the problems occasioned by a zealous commitment to the *localization of mental function* and how *reverse inferences* can generate premature conclusions about the latent contributions of mental processes in a given task. In a similar spirit of self-reflection, a few articles from the past five years have complied a list of related biases, assumptions and problematic practices in neuroimaging (Poldrack et al., 2017; Westlin et al., 2023). While a few of these overlap with those previously discussed, many stem from the inherently human factors that are involved in designing and analysing studies and which have pervasive analogues in other sciences of the mind (e.g. the problem of experimenter bias in social psychology). Poldarck et al. (2017) lump these issues into six broad categories: low power problems, researcher degrees of freedom, multiple comparisons, software errors, insufficient study reporting and the difficulty (and cost) of effective replications in neuroimaging. Westlin et al. (2023) claim that most neuroimaging is plagued by three latent assumptions anchored in a neo-phrenological framework: a localization assumption (which we've reviewed before), a one-to-one mapping assumption and an independence assumption, where neural 'ensembles' are thought to function as discreet units apart from the rest of the brain and embodied organism. Though each of these deserves a thorough review, for the purposes of our task here we'll focus on three related issues: *researcher degrees of freedom*, *low power and small effects* and issues of *circularity and underdetermination*.

Poldrack (2012) notes that the field of fMRI is particularly susceptible to many factors identified by Ioannidis (2005) as 'researcher degrees of freedom', including: its use of small sample sizes, the small effects that are often observed, the number of comparisons made in a typical study, the 'flexibility in designs, definitions, outcomes, and analyses methods' and 'being a "hot" scientific field' (p. 1217). When paired with neuroimagers' initial push – motivated by assumptions about localization – to find the neural realizers of mental functions, one can begin to predict how these prevalent biases might lead to misinterpretations of the results and the subsequent rise of 'blobology'. Probably one of the most memorable cases demonstrating the scope of possible misinterpretations occasioned by this mix of biases and assumptions was demonstrated by Bennett and colleagues (2009) who found significant activation ($p < 0.001$) of voxels in the brain of a *dead* salmon supposedly engaged in an 'open-ended mentalizing task'. With small sample sizes, little replication given the hundreds of dollars an hour it costs to run an fMRI experiment, thousands of possible comparisons thanks to the many voxels canvassed and incomplete methods thanks in part to the novelty of the instrument at hand, it's often possible to easily generate and 'find' spurious results; after all, it's likely that just by sheer chance

some of those voxels will have response profiles that clear the significance hurdle and which might be interpreted as a 'significant' result (consult Poldrack [2012] who demonstrates the ease by which one can find 'significant' activity from a randomly simulated pattern of BOLD data).

Though it should be clear that there are a number of factors at play, low statistical power brought about by small samples and the very nature of the BOLD response – with its relatively minute changes on the order of a few fractions of a percentage change – heighten the risk of type one errors in the results of neuroimaging experiments – that is, where we falsely conclude that some effect did take place (Poldrack, 2018, p. 120). Poldrack does not mince words when he states that 'there are very good reasons to think that a substantial number of findings from neuroimaging research may be false' (2018, p. 120). In turn, to combat this pessimism, we might increase the threshold for significance, or adopt more conservative methods for multiple comparisons, but as we've seen these moves increase the risk of *type two* errors where we reject the plausibly of a true effect (Poldrack, 2012, p. 1217). We could also increase our sample sizes and perform careful replications of prior studies; however, this would geometrically increase the already huge costs involved in neuroimaging, pricing out many younger researchers and concentrating epistemic access and leverage in the hands of a few at wealthy institutions. We could also follow the example of genetics research and make more data more openly accessible, and this is the tactic favoured by Poldrack, who makes a strong case that data sharing and open science practices are some of the few practical tools we have to mitigate this serious crisis in neuroimaging (Poldrack et al., 2017).

Finally we should pause to consider the role of 'circularity' errors and the underdetermination of the BOLD signal, both of which we've seen shades of in prior sections. Circularity or 'non-independence' errors can arise when combing through data that are a priori selected to contain many statistically significant correlations in an effort to prove*, a posteriori*, the statistical significance of a similar correlation (Vul & Pashler, 2012). Poldrack (2018) provides a nice caricature of the inference, which I'll parrot here: Consider that I tell you I 'discovered' that members of a private golf club are on average *more wealthy* than a similarly sized random sample of the population. You'd say 'of course, that's because part of what it is to become a member at a private golf club is the ability to pay exorbitant amounts of money'! So, when a neuroimager claims to have discovered a region sensitive to some task, say looking at pictures of cats, we should examine whether they've committed the same fallacy as, given the low power and multiple comparisons of some fMRI analyses, you're almost always guaranteed to find *some* voxels that meet the threshold of significance (remember our friend, the dead salmon). We've already seen one way around this problem with the advent of MVPA that involves harnessing the power of independent sets of data. We could, for instance, use one task to isolate a ROI and another to measure that region's change in response. Further, we can employ the practice of *preregistration*, whereby we publicly commit to our hypotheses and analyses prior to running our study.

Back to the BOLD signal for a minute, we have to remember that the raw hemodynamic response is a *noisy* signal and one that may not contain 'enough information to unambiguous[ly] resolve which of these [neural] factors have resulted in a measurable signal' (Singh, 2012, p. 1129). That is, there's a risk that the BOLD signal *underdetermines* the neural contributions to a mental process under investigation. This is exacerbated by the choices we make when we model and thus 'clean' the BOLD signal. Westlin et al. (2023) discuss how assumptions (e.g. localist assumptions about neural function) that we use when picking our model of the hemodynamic response can lead to enormous changes in the end results: 'modeling the [hemodynamic function] without presuming its shape. . . . Resulted in a reliable task-based signal increase from ~72% of brain voxels to an average of ~96% of imaged voxels' (p. 3). The lesson to draw here is less that the BOLD response is a bad measurement tool but more that it both requires some level of interpretation *and* that it is not an exhaustive indicator of interesting neural processes that may be underling mentation. To put things differently, a zealous commitment to the BOLD response as *the* measure of neural activity risks constantly committing type two errors: just because we didn't find a response in a given network or neural region, we're not entitled to concluding that that region or network doesn't play some role in a given mental process. Hence, the return of a cautionary nudge towards epistemic humility.

Now we're equipped and ready to review an fMRI study on working memory that I helped design and analyse in order to see some of these trends in action and reflect on those pitfalls that were both present to the experimenters at the time and those which have only been spotted with the benefit of hindsight. My hope here is that by doing this we might be able to better fill in the *human dimension* that is perennially woven into contemporary scientific practice and that we might harness the cumulative weight of these lessons as we move on towards an even more rigorous future practice of neuroimaging.

4 Reflections on *frontoparietal networks involved in categorization and item working memory*

Permit me to set the scene: It's 2011 and I am an NSF *Research Experience for Undergraduates* award recipient who will be spending the summer with Colorado State University's Department of Psychology. During the summer, the ten or so of us were paired with excellent mentors and tasked with designing a study and reporting those results. I was very fortunate to be paired with Kurt Braunlich, then a graduate student, and our PI, Carol Seger, whose lab specializes in neuroimaging subcortical structures (such as the basal ganglia) and identifying the roles they play in supporting learning and other foundational cognitive processes.

We have six weeks to design, run and analyse an fMRI study using a suite of tools, including MATLAB, SPSS, E-Prime, BrainVoyager and SPM, that I – up to that point – had little familiarity with. Needless to say, within a week I started having dreams about my shoddily written code. With that in mind, we decided to maximize our resources by following up on a task-switching project that a previous lab member had spent some time on and which Braunlich was familiar. I was happy with the project as this involved studying working memory, or our capacity to hold information in mind while it's no longer in our environment (Gomez-Lavin, 2021; Gomez-Lavin & Humphreys, 2022), which I had some interest in at the time stemming from my background as a BS in psychology and BA in philosophy double-major.

We set about to determine if we could find a measurable difference in neural activity between two kinds of working memory: working memory for specific item features and working memory for stochastic categories (where the members of each were randomly varied). We were interested in this question as we predicted that both tasks would call upon the same, more general 'cognitive control systems'; however, we also thought that the tasks differed in germane and a priori ways: one requires you to keep specific properties of an object in mind over some delay period, whereas the other requires you to keep categorical information of which set an image belongs to (Braunlich, Gomez-Lavin, & Seger, 2015, p. 146). Intuitively, it might seem easier to keep the category label in mind as opposed to the features of the item, since the category label is an abstraction and allows you to discard all the features specific to an item; then again, the process of abstraction isn't cognitively cheap either. To spoil the ending, we *did* find some interesting differences; in particular, we found different networks which differentially supported what we termed 'item' versus 'category' working memory. In what follows, I'll briefly review the methods and tasks used before turning to the analyses and our findings.

As mentioned, we picked these two tasks since we had a strong hunch (thanks largely to Seger's overwhelmingly thorough and awe-inspiring depth of familiarity with the literature) that we'd find common recruitment in a number of networks involved with the processing of visual information, motor commands and general decision-making processes, while also finding significant differences in networks associated with the 'frontoparietal central executive' involved in managing cognitive resources and task demands (Braunlich et al., 2015, p. 146).

Seventeen participants, the majority of which were members of my fellow NSF-REU cohort, were recruited for the study, during which they had to perform a behavioural version of the task outside the scanner until they achieved 85 per cent proficiency (i.e. 85 per cent 'correct' answers) before spending a further hour in the bore of an MRI machine outside Denver. First, we had participants memorize two 'categories', 'A' and 'B', which were each composed of eight female faces randomly selected from a set of twenty-five. We then familiarized participants with three versions of the task. In each of these, they were first presented with a cue, which you can see in Figure 4.1: 'Match the Specific Face' or 'Match the Category'. If told to 'Match the Specific Face' participants then saw an oval cut-out (matched for size) of a woman's face (all matched for age and race) for 1.5 seconds. Afterwards, the screen went blank and they entered the 'Delay Period' during which we hoped they were actively retaining the stimulus in working memory. Then comes the second stimulus, another face. That face either is identical to the first or different. They are then asked to indicate whether the face was a match or mismatch. In 'Match the Category' conditions participants saw a face and had to recall which category the face belonged to. Our hope was that they would recall this category information during the delay period. After the delay participants were either shown another face and asked if the categories of the two faces matched or not, or they were shown a large letter 'A' or 'B' and asked if the first face matched the category indicated by the letter (Braunlich et al., 2015, p. 149). While in training, participants were given feedback on their responses; however, once in the scanner, participants were not given feedback as they completed two fifteen-minute runs of trials with jittered time intervals between trials (they also completed an anatomical scan and a 'ROI' scan were we attempted to localize some 'standard' regions, such as the fusiform face area). Here's where an important decision comes into play: 'In order to increase the power for analyses . . . both correct and incorrect trials were included in the analyses', and, further, we presented fewer 'Match the Category' trials than 'Match the Specific Face' trials (Braunlich et al., 2015, p. 149). Worsening our low power situation, we also excluded one participants' data for excessive head movement (Braunlich et al., 2015, p. 149).

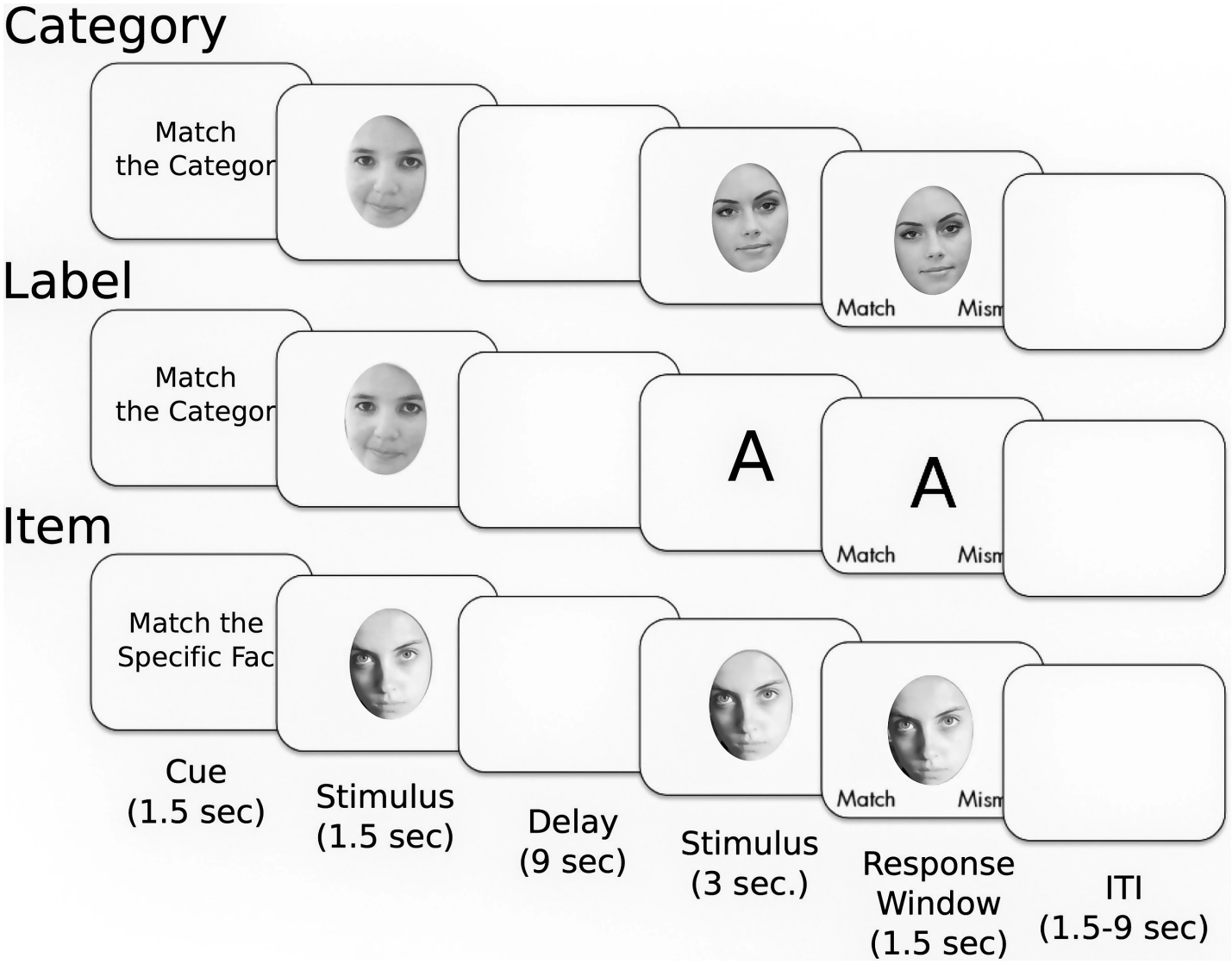

Figure 4.1 A depiction of the stimuli and three task trials that participants completed as part of our study. Participants were trained to 85 per cent task proficiency before scanning, and ITIs were jittered during scanning. Reproduced from *NeuroImage,* vol. 107, Braunlich et al., 'Frontoparietal networks involved in categorization and item working memory'. © 2015, with permission from Elsevier. Faces in image © 2023 Free-images.com Pixabay. Public Domain

We then took the BOLD response data from our participants and performed two kinds of analyses: a univariate GLM that incorporated a 'canonical' model of the hemodynamic response *and* a constrained principle components analysis that used a 'finite-impulse response' model of the hemodynamic response that could be tuned to different profiles that might be expected across a variety of functional networks (Braunlich et al., 2015, p. 149). For the univariate analyses, the timings of the epoch (e.g. when stimuli were presented and when participants were asked for their decisions) were used as independent 'boxcar' regressors in a whole-brain analysis of activity so that for instance the encoding regressor was modelled as the 1.5 seconds of activity around the presentation of the first face stimulus to be remembered (Braunlich et al., 2015). We found significant differences during encoding, delay period and response 'probe', some of which are depicted in Figure 4.2. Possibly the most striking contrast is seen in the bottom row of Figure 4.2, where participant task demands were most similar. Here we saw that 'category trials elicited greater activity than Item trials in executive regions of the cerebellum, frontal (middle frontal, anterior insula/inferior frontal, and superior medial gyrus) and parietal regions (inferior parietal, angular gyrus, and precuneus), including the salience network', which fell in line with our initial predictions (Braunlich et al., 2015, p. 151).

Constrained principle component analyses were also used to determine which voxels were responsible for most of the variance within the raw BOLD signal, which led us to, in turn, five principle components that contained most of this variance (Braunlich et al., 2015, p. 149). To do this we regressed a large matrix containing all of the BOLD data onto a model of the 'finite-impulse response', which generated a large series of weights that contained the variance in the initial colossal BOLD matrix, allowing us to extract the principle components which were then processed and overlaid on the anatomical models as can be seen in Figure 4.3.

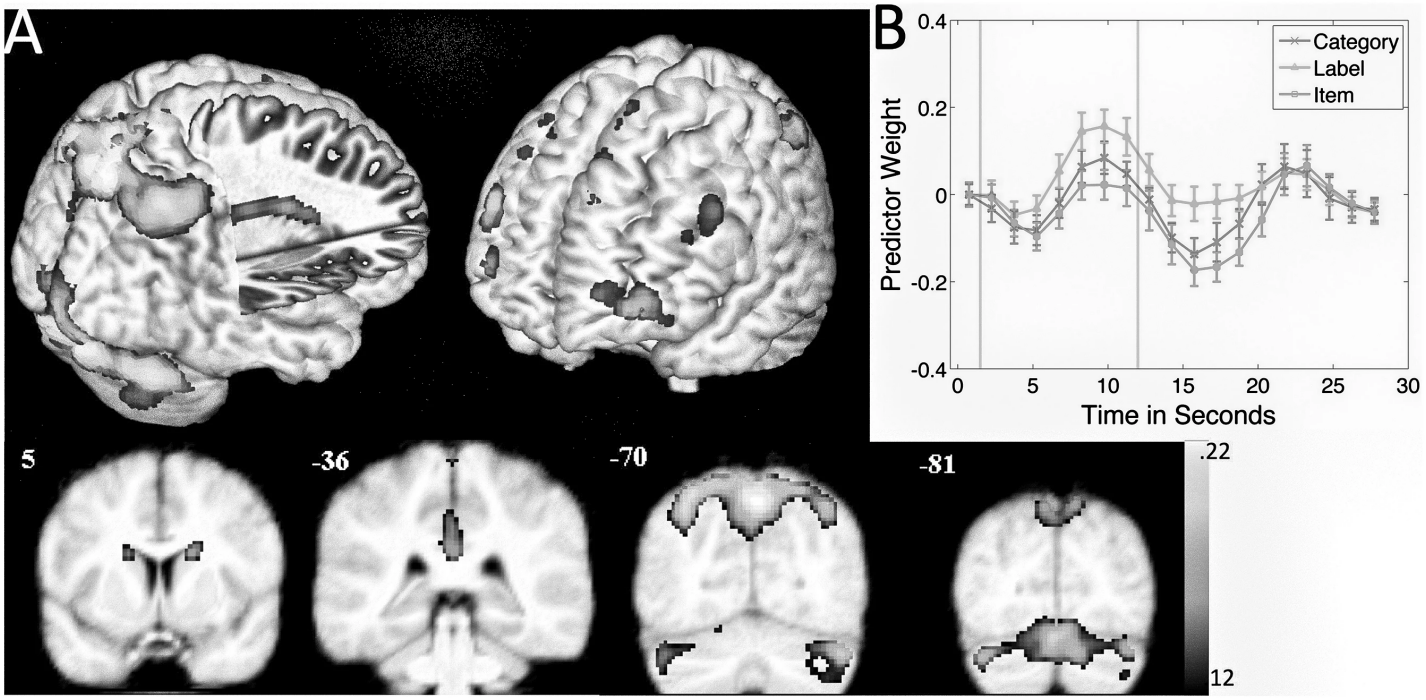

Figure 4.3 'Component 3. Note the recruitment of FP-CEN regions including the lateral prefrontal cortex and intraparietal sulcus, along with the cerebellum and caudate. (A) The top 5% of component loadings overlaid on the MNI template provided by MRIcron (3d renderings, top) and an averaged structural image (slices, bottom). (B) Predictor weight timecourse. Error bars represent the standard error of the mean. Vertical lines indicate onsets of visual stimuli' (p. 153). Reproduced from *NeuroImage,* vol. 107, Braunlich et al., 'Frontoparietal networks involved in categorization and item working memory'. © 2015, with permission from Elsevier

At first glance, one might be tempted to interpret the bright orange blobs in Figure 4.3 as hotspots of greater neural activity, much in the same way as you might be licensed to do so with the spots in Figure 4.2. However, and as we cautioned readers in the original article, that would be a mistake since the assumptions driving the two kinds of analyses fundamentally differ, and this entails a difference in what is represented by the statistical maps drawn as 'blobs' in each of the respective images (Braunlich et al., 2015, p. 151). Rather, the spots in Figure 4.3 hint at how different networks and regions are variously recruited depending on the task demands that the subject faces. Much in the same way that MVPA helps us capture a pattern of activation specific to a given stimulus, CPCA helps us depict a pattern of task-dependent variance across voxels, and hence networks. Finally, we combined our two analyses methods to yield a 'CPCA-masked univariate' analysis, which has some significant limitations given the radically different assumptions of each of these methods – including, for instance, the different models of the hemodynamic response and how each treats variance (e.g. where this variance is usually discarded in univariate analyses) (Braunlich et al., 2015, p. 153). These further masked analyses then allowed us to observe some exploratory connections between regions recruited for different events in our task and their functional roles; for instance, we found that the fusiform was more active when participants were asked to retain a face as opposed to a category, from which we concluded that 'these visual regions have both feature specific processing roles, and functional roles within the task-related salience network in responding to stimuli', as they seemed to modulate their activity based on the task demands (Braunlich et al., 2015, p. 154).

5 Pitfalls and future aspirations towards a more accessible and open science of fMRI

Overall, I look back on this project with a glint of pride, given that it was borne from a six-week pilot study largely patched together by a mathematically disinclined undergraduate with the help of an amazingly patient mentor who went on to run the task on additional subjects and perform the non-univariate analyses. Despite some hiccups and issues which can be traced to the epistemic pitfalls we detailed earlier, my hope and expectation is that the general findings approach their mark, namely that distinct functional networks are involved in these two kinds of cognitive operations, categorization and specific item retention, and that we managed to glimpse some of the structure of these networks.

However before concluding, we have to do the difficult work of picking out just where we ran into *epistemic* trouble and how we might best avoid these pitfalls in a future study. Roughly, and with a large dose of humility, we can divvy up the pitfalls into two camps: those involving our methodical and procedural choices and those involving our analyses and conclusions.

The methodological issues, which you might have already begun to pick out, revolve around our use of small sample sizes, the particularities of that sample and, finally, the nature of the task and stimuli choice. We only used seventeen participants, one of which was excluded from the analyses due to excessive movement within the scanner. Stop for a second and imagine being placed in a small bore of a huge, loud and uncomfortable superconducting magnet while being instructed not to move your head more than 3 mm or else that trial's worth of data will have to be tossed. And now imagine doing this *for an hour*. It's surprising that the cost of this kind of labour ($20) isn't a larger fraction of a study's expense! So who willingly subjects themselves to this kind of experience? By in large, we had the fortune of easy access to a pool of driven undergraduates who were members of this NSF-REU cohort. They make excellent volunteers, indeed most trained to 85 per cent proficiency on the task in only a handful of runs; however, it's simultaneously a strange sample to be drawing conclusions from that apply to the general population. This points to our next problem: the ease of the task and the choice of stimuli. Why only twenty-five faces? Why only faces of white women? Why only eight faces per category? Why faces at all? Answers to these questions largely boil down to two factors: we (really, I) had six weeks to design a pilot study and obtain fMRI data, and another lab in the department had ready access to these images and they were willing to share them. It's remarkable how many downstream consequences come from the practical considerations that filter into a study's initial design: free stimuli, ready volunteers, a rookie coder, all these factors add up.

If I could rerun the study, it would be ideal to have a larger and more diverse sample, and it would be ideal to feature conditions that were more cognitively taxing and which showcased other non-face-related stimuli (maybe even mixing and matching between stimulus types in the categorization condition). However, much of that comes at a material cost of time and money. As such, in terms of *future*-directed lessons that we could draw from this, besides having more time and money, it's clear that we could have benefited from openly accessible stimuli and task libraries (which are now more common with tools like PsychoPy) and openly accessible data from other, similar fMRI studies. These are points elaborated at length by Poldrack (2018), but it's clear that one of the few ways to make this technology more accessible to greater numbers of junior and diverse scientists is by making more data and tools accessible. In some important sense, tackling the issue of accessibility lies at the heart of solving many of the epistemic issues surrounding fMRI.

There are also apparent issues with the analyses and interpretations of our results, most of which we readily admitted were limitations to the study. Because of our small sample and overall low statistical power, we had to include runs where the participants were mistaken (Braunlich et al., 2015, p. 149). Our effects, including most of our condition-sensitive main effects captured by the five components we extracted, were rather small, with many effects in the low single-digit percentage values of eta-squared, indicating a small effect size (cf. the eta-squared values for Component 3, depicted in Figure 4.3 and found on page 152 of the original paper). Furthermore, the paper uses a mixture of analytical methods, moving between univariate and constrained principle components analyses and even mixing between them. We acknowledge these limitations, but again this choice was borne of the pragmatic considerations that existed at the start of the study. Further, mixing designs arguably makes the results harder to interpret and more subject to 'exploratory' takes that may not have preceded the collection of data, although we also admit as much when we move into more exploratory discussions of our results. It's not that exploratory results are bad *simpliciter*, but without a record of predictions and analyses that precede data collection – as is now becoming standard with *preregistrations* and *registered reports* – it can become difficult to parse what is a test of a hypothesis and what is an exploratory analysis. Ideally, the evidential weight of results should reflect their epistemic provenance, and preregistrations – although cumbersome – are good tools to help us become better stewards of our data. Finally, in the paper we engage in some light reverse inference, for instance, when we infer the 'rapid allocation of cognitive resources' managed by the salience network in some conditions (those requiring faces) because of a distinct pattern of activation in the insula and anterior cingulate (Braunlich et al., 2015, p. 154). These are usually illicit moves; however, given the nature of the tasks, the large literature on offer and the many independent and combined analyses, such an inference is, if not fully justified, not without *some* warrant. Were it the only reported result, then that might be spurious, but given that it is but one among many, one need only modify their credence in its truth slightly. As I've hinted, I believe many of these issues can be mitigated by the adoption of the practice of preregistration and other open science principles that curb some of these impulses and allow for more transparency in the scientific process.

This chapter has reviewed the physical basis of fMRI and provides a survey of its history, methods and analyses. In doing so, we've precipitated a number of *epistemic pitfalls* that can emerge from the selection, design and execution of an fMRI study. I then provided a review of a study that I was involved in carrying out, and we analysed the pitfalls involved in that study. From this process we've generated a few overall lessons about the value of epistemic humility in neuroscientific research and of the great value and immediate need for the adoption of more open science principles that could help to maintain fMRI's privileged place as a tool of choice for non-invasive investigation of the neural dynamics and properties of the brain for decades to come.

References

Bandettini, P. A. (2012). Twenty years of functional MRI: The science and the stories. *NeuroImage*, *62*(2), 575–588. https://doi.org/10.1016/j.neuroimage.2012.04.026

Bennett, C., Miller, M. & Wolford, G. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *NeuroImage*, *47*, S125. https://doi.org/10.1016/S1053-8119(09)71202-9

Biswal, B. B. (2012). Resting state fMRI: A personal history. *NeuroImage*, *62*(2), 938–944. https://doi.org/10.1016/j.neuroimage.2012.01.090

Braunlich, K., Gomez-Lavin, J. & Seger, C. A. (2015). Frontoparietal networks involved in categorization and item working memory. *NeuroImage*, *107*, 146–162.

Buxton, R. B. (2012). Dynamic models of BOLD contrast. *NeuroImage*, *62*(2), 953–961. https://doi.org/10.1016/j.neuroimage.2012.01.012

Clark, V. P. (2012). A history of randomized task designs in fMRI. *NeuroImage*, *62*(2), 1190–1194. https://doi.org/10.1016/j.neuroimage.2012.01.010

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? (Position Paper Presented to the European Cognitive Neuropsychology Workshop, Bressanone, 2005). *Cortex*, *42*(3), 323–331. https://doi.org/10.1016/S0010-9452(08)70358-7

Courtney, S. M. (2012). Development of orthogonal task designs in fMRI studies of higher cognition: The NIMH experience. *NeuroImage*, *62*(2), 1185–1189. https://doi.org/10.1016/j.neuroimage.2012.01.007

Dennet, D. (1991). *Consciousness explained*. New York: Little, Brown, and Co.

Fodor, J. A. (1999). Diary: Why the brain. *London Review of Books*, *21*(30). https://www.lrb.co.uk/the-paper/v21/n19/jerry-fodor/diary

Formisano, E. & Kriegeskorte, N. (2012). Seeing patterns through the hemodynamic veil—The future of pattern-information fMRI. *NeuroImage*, *62*(2), 1249–1256. https://doi.org/10.1016/j.neuroimage.2012.02.078

Glymour, C. & Hanson, C. (2016). Reverse inference in neuropsychology. *The British Journal for the Philosophy of Science*, *67*(4), 1139–1153. https://doi.org/10.1093/bjps/axv019

Goebel, R. (2012). BrainVoyager—Past, present, future. *NeuroImage*, *62*(2), 748–756. https://doi.org/10.1016/j.neuroimage.2012.01.083

Gomez-Lavin, J. (2021). Working memory is not a natural kind and cannot explain central cognition. *Review of Philosophy and Psychology*, *12*, 199–225.

Gomez-Lavin, J. & Humphreys, J. (2022). Striking at the Heart of Cognition: Aristotelian Phantasia, working memory, and psychological explanation. *Medicina Nei Secoli: Journal of History of Medicine and Medical Humanities*, *34*(2), 13–38.

Hardcastle, V. G. & Stewart, C. M. (2002). What do brain data really show?*Philosophy of Science*, *69*(S3), S72–S82. https://doi.org/10.1086/341769

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, *62*(2), 852–855. https://doi.org/10.1016/j.neuroimage.2012.03.016

Huettel, S. A. (2012). Event-related fMRI in cognition. *NeuroImage*, *62*(2), 1152–1156. https://doi.org/10.1016/j.neuroimage.2011.08.113

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124. Epub 2005, August 30. Erratum in: *PLoS Med*, 2022, August 25, *19*(8), e1004085. PMID: 16060722; PMCID: PMC1182327.

Klein, C. (2010). Images are not the evidence in neuroimaging. *The British Journal for the Philosophy of Science*, *61*(2), 265–278. https://doi.org/10.1093/bjps/axp035

Klein, C. (2014). The brain at rest: What it is doing and why that matters. *Philosophy of Science*, *81*(5), 974–985. https://doi.org/10.1086/677692

Kwong, K. K. (2012). Record of a single fMRI experiment in May of 1991. *NeuroImage*, *62*(2), 610–612. https://doi.org/10.1016/j.neuroimage.2011.07.089

Lv, H., Wang, Z., Tong, E., Williams, L. M., Zaharchuk, G., Zeineh, M., Goldstein-Piekarski, A. N., Ball, T. M., Liao, C. & Wintermark, M. (2018). Resting-state functional MRI: Everything that nonexperts have always wanted to know. *American Journal of Neuroradiology*, ajnr;ajnr.A5527v1. https://doi.org/10.3174/ajnr.A5527

Machery, E. (2014). In defense of reverse inference. *The British Journal for the Philosophy of Science*, *65*(2), 251–267. https://doi.org/10.1093/bjps/axs044

McCaffrey, J. & Danks, D. (2022). Mixtures and psychological inference with resting state fMRI. *The British Journal for the Philosophy of Science*, *73*(3), 583–611. https://doi.org/10.1093/bjps/axx053

Ogawa, S. (2012). Finding the BOLD effect in brain images. *NeuroImage*, *62*(2), 608–609. https://doi.org/10.1016/j.neuroimage.2012.01.091

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

Poldrack, R. (2018). *The new mind readers: What neuroimaging can and cannot reveal about our thoughts*. Princeton: Princeton University Press.

Poldrack, R. A. (2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed?*Perspectives on Psychological Science*, *5*(6), 753–761. https://doi.org/10.1177/1745691610388777

Poldrack, R. A. (2012). The future of fMRI in cognitive neuroscience. *NeuroImage*, *62*(2), 1216–1220. https://doi.org/10.1016/j.neuroimage.2011.08.007

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E. & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, *38*(1), 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030

Roskies, A. L. (2007). Are neuroimages like photographs of the brain?*Philosophy of Science*, *74*(5), 860–872. https://doi.org/10.1086/525627

Singh, K. D. (2012). Which 'neural activity' do you mean? FMRI, MEG, oscillations and neurotransmitters. *NeuroImage*, *62*(2), 1121–1130. https://doi.org/10.1016/j.neuroimage.2012.01.028

Snyder, A. Z. & Raichle, M. E. (2012). A brief history of the resting state: The Washington University perspective. *NeuroImage*, *62*(2), 902–910. https://doi.org/10.1016/j.neuroimage.2012.01.044

Takagi, Y. & Nishimoto, S. (2022). High-resolution image reconstruction with latent diffusion models from human brain activity [Preprint]. *Neuroscience*. https://doi.org/10.1101/2022.11.18.517004

Uttal, W. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.

Vul, E. & Pashler, H. (2012). Voodoo and circularity errors. *NeuroImage*, *62*(2), 945–948. https://doi.org/10.1016/j.neuroimage.2012.01.027

Westlin, C., Theriault, J. E., Katsumi, Y., Nieto-Castanon, A., Kucyi, A., Ruf, S. F., Brown, S. M., Pavel, M., Erdogmus, D., Brooks, D. H., Quigley, K. S., Whitfield-Gabrieli, S. & Barrett, L. F. (2023). Improving the study of brain-behavior relationships by revisiting basic assumptions. *Trends in Cognitive Sciences*, S1364661322003321. https://doi.org/10.1016/j.tics.2022.12.015

Notes

[1] However, for those brave enough and interested the Abler Einstein College of Medicine has made a fifty-six-part lecture series on MRI available via their YouTube channel: https://www.youtube.com/watch?v=35gfOtjRcic&ab_channel=AlbertEinsteinCollegeofMedicine.

[2] Indeed, for a thorough review of the technical antecedents that contributed to a number of aspects of fMRI, please consult the 2012 special issue of *NeuroImage* which collected nearly 100 articles from contributors central to the foundation, early development and use of fMRI over its first twenty years (Bandettini, 2012). Much of what I write in this section is indebted to a review of this thorough collection.

[3] Most medical and research scanners operate at 3T or Telsa, with the average magnetic field exerted by the earth at its surface coming in the order of microteslas.

[4] Note the common refrain along the line that while the brain only takes up 2 per cent of a typical human's weight it uses something like 20 per cent of the overall energy of the body (Snyder & Raichle, 2012, p. 904).