Begging the Question: The Non-Independence Error in fMRI Data Analysis

Edward Vul, Nancy Kanwisher

You sit with a deck of fifty-two cards in front of you, face down.
You flip the first card: a 10 of diamonds.
What are the odds of that? One out of fifty-two.
Remarkable.
You flip the next card: a queen of hearts.
Unbelievable! The odds of this sequence were 1/52 * 1/51 (less than 1 in 2000).
You continue flipping cards: a 4 of clubs follows, then an 8 of diamonds, then an 8 of hearts. Once you have flipped them all over you stare in disbelief; the particular sequence of cards you just observed happens one out of every $8*10^{67}$ (52 factorial) times. Every person in the world could shuffle a deck of cards and flip through it every minute of their entire lives, and even then, the odds of the world seeing your particular sequence of cards will be less than $1/10^{50}$!
Extraordinary.

Something is very wrong here. The conclusion is absurd. Yet similar logic is prevalent in both lay and scientific reasoning. Some have used variants of this argument to account for the origins of humans on earth: Proteins could be randomly shuffled for eons before humans emerged in all their glory. Since the likelihood of human existence by pure chance is so slim, surely intelligent design is the most parsimonious explanation. The card example was introduced to illustrate just how preposterous this objection to evolution is. Unfortunately, this logical fallacy, which we will call here the "non-independence" error, is not restricted to arguments from the scientifically unsophisticated. It is prevalent in cognitive neuroscience as well. For instance, of the eight papers in a recent special issue of *Neuroimage*, five contained variants of this error[1]. The prevalence of this error is troubling because it can produce apparently significant effects out of pure noise (Figure 1). In this chapter we will describe the error formally, consider why it appears to be more common in fMRI than other fields, provide examples of this error in its most common guises, and propose a few heuristics that may help lay people and scientists alike avoid the error.

1. Formal description of the non-independence error

What exactly is the error that leads to the absurd conclusion in the card example? We can describe it in different theoretical frameworks: statistical hypothesis testing, propositional logic, probability theory, and information theory. These frameworks are rarely discussed together, and never connected in the context of the non-independence error. In this section we describe the error in the context of statistical hypothesis testing; in the Appendix we consider it from the three other perspectives.

In statistical hypothesis testing, the most common non-independence error is referred to as 'selection bias'. Essentially all statistical models used for hypothesis testing assume that the sampling (selection) process is independent of the relevant measure. 'Selection bias' is a violation of this independence assumption.

[1] (den Ouden, Frith, Frith, & Blakemore, 2005; Gillath, Bunge, Shaver, Wendelken, & Mikulincer, 2005; Harris, Todorov, & Fiske, 2005; Mitchell, Banaji, & Macrae, 2005; Sander et al., 2005)

If we assume that our deck of cards is a random sample from the population of all decks of cards, and we are evaluating the likelihood that such a deck of cards will have a particular order specified in advance, we would be surprised to find such a coincidence (indeed, $p<10^{-67}$). However, our sampling process is very different. Our sample was not drawn from the population of all random decks; instead, it was a sample from all decks that we just observed to have the particular order in question.

Statistics textbooks often describe "selection bias" via simplistic examples of samples that are not representative of the population (e.g., drawing data exclusively from the NBA when assessing the height of Americans). Obviously, in such cases, the sample will clearly be different from the population, and generalizations to the population would be unjustified because they rely on the assumption that the sampling process is independent of the measure. This classical form of selection bias effectively conveys the intuition that a sample ought to be representative of the population. However, 'representative of the population' is loosely defined. If we seek to test whether a particular group (say, people who are taller than 6' 5") somehow differs from the population (say, have higher salaries), then, if such an effect truly exists, the process of 'selecting' our population of interest necessarily gives us a sample different from the global population, and there is nothing wrong in this case.

We can define 'selection bias' more formally than 'not representative of the population' as follows: if our selection criteria are applied to a sample from the null hypothesis distribution, the selected subset of that sample must also be a sample from the null hypothesis. For example, if we seek to evaluate whether a sociology class is more difficult than a psychology class, we might find a sample of students who have taken both, and evaluate the difference in this group's average score in the two classes. Let's imagine that the average grades for sociology and psychology are identical in this sample (thus, we have a sample from the null hypothesis distribution -- there is no effect in the sample). Now imagine that from this sample, we choose only students who had a better grade in sociology than psychology. This newly selected subsample will have a higher average grade in sociology than psychology. So our analysis procedure "(1) select all students with a higher grade in sociology than psychology, (2) evaluate average scores for sociology and psychology in this sample" violates the requirement that the selection criteria not alter the null hypothesis distribution when applied to it. Thus our selection is biased.

This definition of selection bias can be expressed in terms of independence in probability theory: if X is a random variable representing our data, and P(X) reflects the probability distribution assumed by the null hypothesis, then P(X|C) where C is the selection criteria, must be equal to P(X), the null hypothesis distribution. Thus, 'selection bias' is a violation of independence between selection and the subsequent statistical measure. Though this point may appear obvious or trivial, it is crucial when considering examples further removed from our intuitions about circular reasoning or population representativeness.

2. Examples of the non-independence error in fMRI

The non-independence error arises in fMRI when a subset of voxels is selected for a subsequent analysis, but the null-hypothesis of the analysis is not independent of the selection criteria used to choose the voxels in the first place. Take the simplest practical case: If one selects only voxels in which condition A produces a greater signal change than condition B, and then evaluates whether the signal change for conditions A and B differ in those voxels using the

same data, the second analysis is not independent of the selection criteria. The outcome of this non-independent second analysis is *statistically guaranteed* and thus *uninformative:* A will be greater than B, since this conclusion is presumed by the selection criterion (Culham, 2006). Furthermore, this outcome will be *biased*: given that the data will be perturbed by random noise, selecting voxels in which A>B preferentially selects voxels in which the random noise is positive for A and negative for B.

There are many ways for the combination of voxel selection and subsequent analysis to produce a non-independence error and thus biased results. However, neither a particular selection method nor a particular analysis method is alone sufficient for a violation of independence; the violation results from the relationship between the two. We will describe a few situations in which the particular combination of selection method and subsequent analysis results in non-independence. We start with the simplest cases, where the error will be most obvious and go on to more elaborate cases where non-independence is harder to spot.

fMRI analyses that contain the non-independence error often jeopardize the conclusions of the study for two reasons. First, the non-independent analysis will be statistically biased, rendering its conclusions limited (at best) and completely invalid (at worst). Second, because researchers often rely on the secondary analysis to support their claims, the initial analysis used to select voxels is often not statistically sound because it is not properly corrected for multiple-comparisons. Insufficient correction for multiple-comparisons guarantees that some number of voxels will pass the statistical threshold, and offers no guarantee that they did so because of some true underlying effect rather than fortuitous noise. In cases when both a lax multiple-comparisons correction is employed and a non-independent secondary test is used, all of the conclusions of the experiment are questionable: the lax selection threshold may have only selected random fluctuations, and a seemingly significant result may have been produced literally out of noise (Figure 1, reprinted from (Baker, Hutchison, & Kanwisher, 2007)).

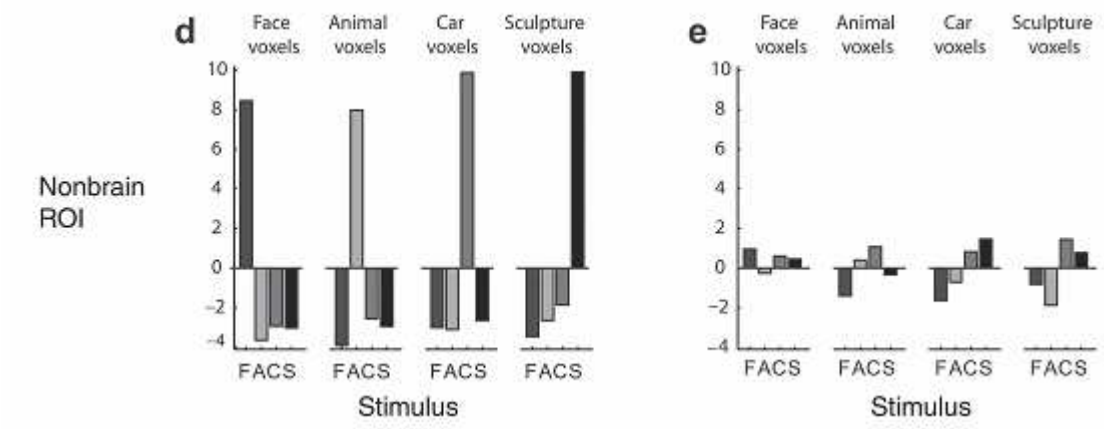

Figure 1. A portion of the graph from Baker et al (2007). (d) A non-independent analysis and (e) the suggested independent analysis were performed on non-brain fMRI data (the nose). With a non-independent analysis, even the nose may be shown to contain face-, animal-, car-, and sculpture- selective voxels. A sound, independent analysis does not produce such spurious results.

2.1 Testing the signal change in voxels selected for signal change

The most damaging non-independence errors arise when researchers perform statistical tests on non-independent data. In these cases, the non-independent tests are explicitly used in support of a conclusion, while the non-independence renders the test biased, uninformative, and invalid.

One such example can be seen in a recent paper by Summerfield and colleagues (Summerfield et al., 2006). The authors sought to identify category-specific predictive codes. Subjects were exposed to "face set" blocks in which they had to identify faces, and "house set" blocks in which, for identical stimuli, subjects had to identify houses. In Figure 3a Summerfield et al show the voxels that are more active for face set blocks than house set blocks (selected at a low, uncorrected threshold: $p<0.01$). The authors then take the maximally significant voxel in this contrast and run "post hoc ANOVAs". The ANOVAs are defined with two factors: one factor is face set versus house set, and the second factor is stimulus (which *is* independent of face or house set because different stimuli types were equally distributed in the two set conditions). The result of this ANOVA is a significant main effect of face set, with remarkably high F statistics reported.

However, these ANOVAs were run only on the maximally active voxel in the face-set minus house-set contrast, defined by the same data. This means that the voxels are guaranteed to have a significant main effect of face-set greater than house-set. The statistics reported for the main effect in the ANOVA add no new information, and do not bolster the authors' claims in the slightest. The ANOVA results were presupposed by the selection criteria.

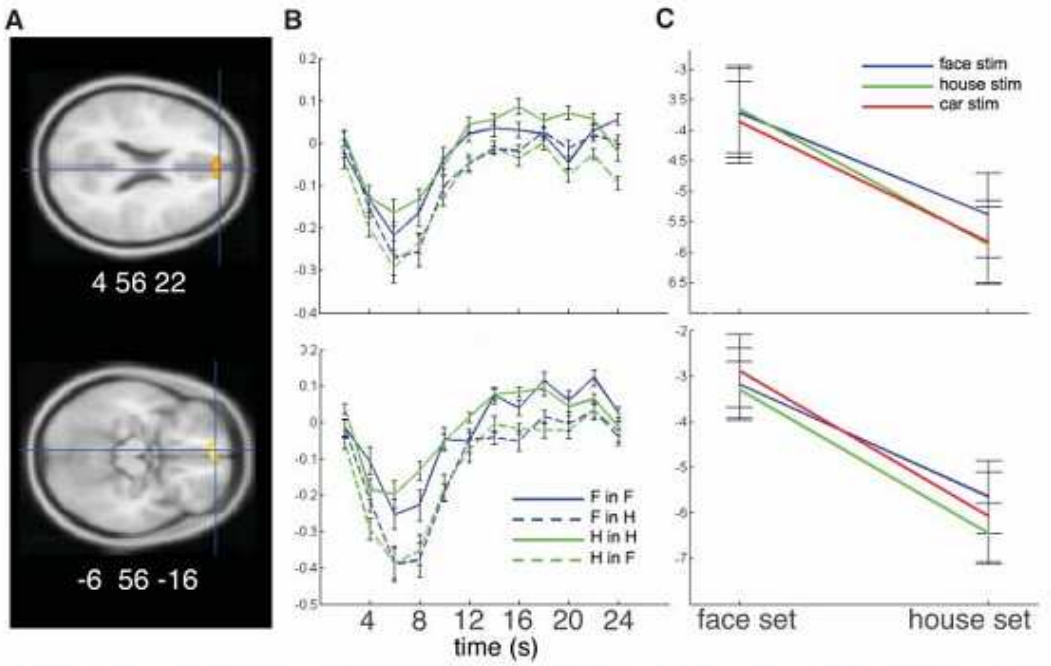

Fig. 3. dMFC and vMFC respond to face set. (**A**) Statistic parametric maps showing dMFC (top) and vMFC (bottom) voxels responding to face set blocks > house set blocks, rendered at a statistical threshold of $P < 0.01$. (**B**) Evoked hemodynamic responses, as in Fig. 2. Continuous lines are face set trials; dashed lines are house set trials. (**C**) Post hoc ANOVAs at the peak voxel in each cluster revealed a significant main effect of set [dMFC: 4, 56, 22; $F_{(2,14)} = 22.1$, $P < 0.0004$; vMFC: −6, 56, −16; $F_{(2,14)} = 20.4$, $P < 0.0005$]. No effect of stimulus was observed at either dorsal ($P = 0.70$) or ventral ($P = 0.75$) sites.

Figure 2. Voxels selected for having a main effect of set (at a low, uncorrected threshold) are reported as having a very significant main effect of set. Reprinted from (Summerfield et al., 2006).

These statistics are misleading. The ANOVA results are used to bolster the whole brain analysis that identified the regions. The whole brain analysis itself used a particularly low threshold ($p<0.01$, without multiple comparisons correction), and as such could not stand on its own. However, it effectively imposes a bias on the subsequent ANOVA. It is quite possible that the results displayed in Figure 3 may be entirely spurious: the results of the whole-brain analysis may be due to chance (false positives), and the results of the ANOVA are guaranteed given the selection criteria. This is exactly the sort of analysis that has motivated us to write this chapter: the numbers reported (F values greater than 20) appear convincing, but they are meaningless.

Non-independent statistical tests appear in numerous other high-profile articles, e.g. (Cantlon, Brannon, Carter, & Pelphrey, 2006; Grill-Spector, Sayres, & Ress, 2006; Piazza, Pinel, Le Bihan, & Dehaene, 2007; Ruff et al., 2006; Todd & Marois, 2004). Although non-independent tests are prevalent in the literature, the conclusions of an entire paper do not always depend on that biased analysis. However, in some cases (Summerfield et al., 2006) the researchers may have produced their main significant result out of nothing.

2.2 Plotting the signal change in voxels selected for signal change.

The most common, most simple, and most innocuous instance of non-independence occurs when researchers simply plot (rather than test) the signal change in a set of voxels that were selected based on that same signal change.

2.2.1 *Selecting an interaction*

Take for instance a classic article about the effect of load on motion processing (Rees, Frith, & Lavie, 1997). In this study, the researchers sought to test Lavie's load theory of attention – that ignored stimuli will be processed to a greater degree under low-load compared to high-load attended tasks. Subjects performed either a difficult or an easy linguistic task at fixation (to make load high or low, respectively), while an ignored dot field either moved or did not move in the background. The authors predicted a greater difference in activation between motion and no motion conditions during low load compared to high load. Thus, Rees et al. found the peak voxel within some distance of area MT in which this interaction was greatest, and plotted the signal change in that voxel.

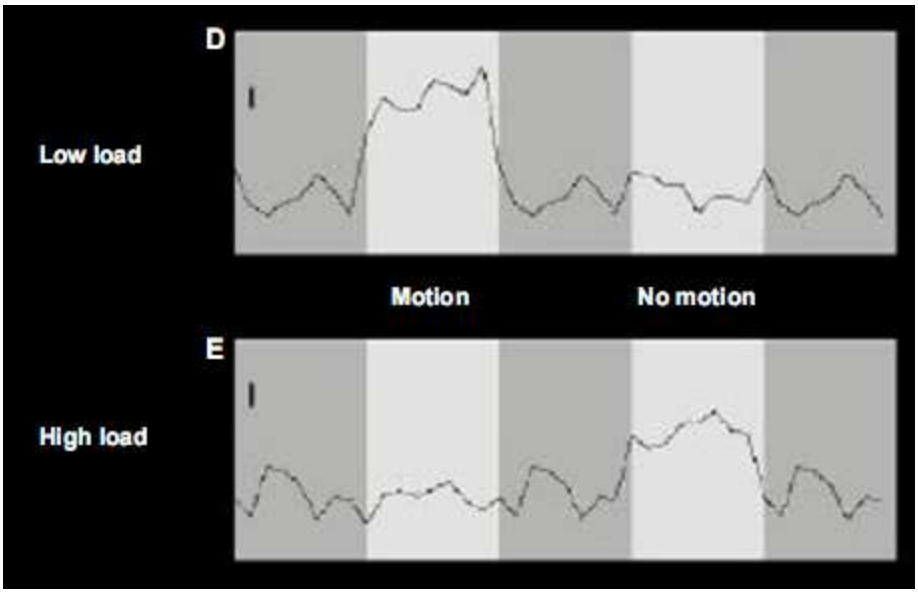

Figure 3. Time course from the peak voxel selected for an interaction is biased towards implausible activation patterns (no motion > motion in MT; under high load). Reprinted from (Rees et al., 1997)

This graph appears informative of the hypothesis under investigation, but it is not: the presence of the predicted interaction was a pre-requisite for data to be included in this graph. Of all voxels that could be processing motion, only the one with the most significant interaction is displayed. So, of course an interaction will be apparent in this plot. It is, however, the case that this graph may be used to evaluate other, orthogonal aspects of the data. We will discuss this in section 2.2.3. For now, it is important to note that the average activity of voxels that were selected under some hypothesis is not diagnostic of the hypothesis, and should never be used as implicit evidence.

This graph of signal change in the peak voxel selected for an interaction is also an excellent example of how non-independence introduces bias. The hypothesis used for selection was that the difference (Motion - No Motion) will be greater under low load than high load. Imagine that the ground-truth about V5 (the area being tested in this experiment) is what is predicted by Lavie's theory: under low load, motion produces greater activation than no motion, but under high load, there is no difference. The measured activity in each voxel will be perturbed by noise around this ground truth. If we then select only the voxels for which LowLoad(Motion-NoMotion) is much greater than HighLoad(Motion-NoMotion), we will be preferentially selecting voxels in which the noise that corrupts ground truth causes Motion-NoMotion to be *negative* under high load conditions (as this will maximize the interaction). This is precisely what we see in the data that were selected by this interaction test – under high-load, MT, a motion-processing region, is more active under the No Motion condition than the Motion condition. This otherwise unintelligible result perfectly illustrates the effects of noise selection, that is, how the selection criteria favor certain patterns of noise over others.

2.2.2 Selecting an effect

Although plotting data that were chosen because they contained an interaction provides an excellent example of noise selection, it is much more common for main effects to be selected. Take for instance a recent article by (Thompson, Hardee, Panayiotou, Crewther, & Puce, 2007).

In this study the researchers sought to identify regions that respond to perceived hand motion, and perceived face motion. Subjects were scanned as they viewed blocks of fixation (baseline), hand motion, face motion, and radial grating motion. Regions of interest (ROIs) were selected based on the following two criteria: (a) the average response in these regions to stimulus blocks was greater than the average response to baseline (fixation) blocks; and (b) these regions showed greater activation for hand or face motion compared to radial grating motion. Figure 2 (p. 969) shows the ROIs that were selected as exhibiting face-motion preference, hand-motion preference, and both hand and face motion preference. Below each ROI, the authors plot the percent signal change within that ROI for each of the three conditions (face, hand, and radial motion) relative to fixation.

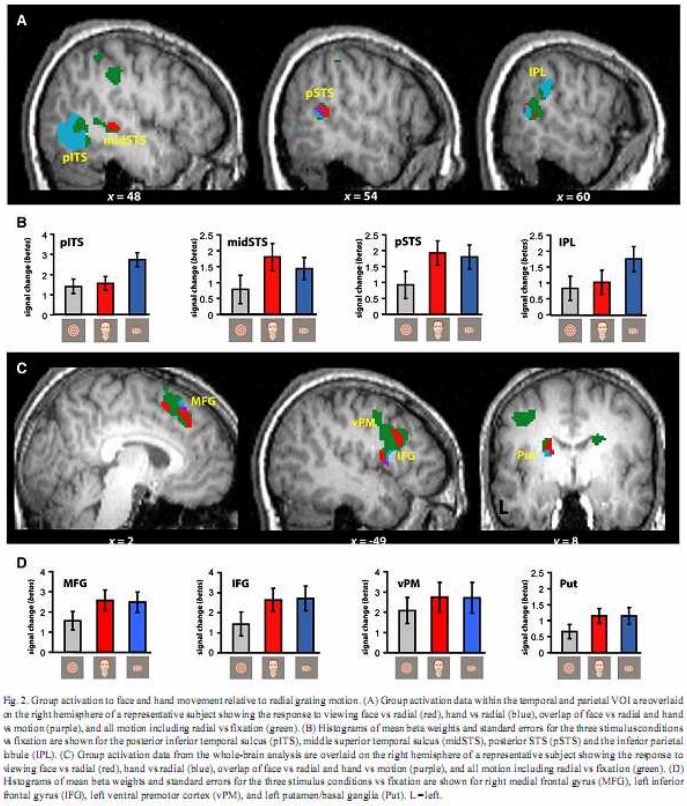

Fig. 2. Group activation to face and hand movement relative to radial grating motion. (A) Group activation data within the temporal and parietal VOI are overlaid on the right hemisphere of a representative subject showing the response to viewing face vs radial (red), hand vs radial (blue), overlap of face vs radial and hand vs motion (purple), and all motion including radial vs fixation (green). (B) Histograms of mean beta weights and standard errors for the three stimulus conditions vs fixation are shown for the posterior inferior temporal sulcus (pITS), middle superior temporal sulcus (midSTS), posterior STS (pSTS) and the inferior parietal lobule (IPL). (C) Group activation data from the whole-brain analysis are overlaid on the right hemisphere of a representative subject showing the response to viewing face vs radial (red), hand vs radial (blue), overlap of face vs radial and hand vs motion (purple), and all motion including radial vs fixation (green). (D) Histograms of mean beta weights and standard errors for the three stimulus conditions vs fixation are shown for right medial frontal gyrus (MFG), left inferior frontal gyrus (IFG), left ventral premotor cortex (vPM), and left putamen/basal ganglia (Put). L = left.

Figure 4. The percent signal change in areas selected for each of those percent signal change profiles. Despite the error bars, these bar-graphs do not contain any information about the main claims of the paper. Reprinted from (Thompson et al., 2007).

What exactly could one learn from these plots and what aspects of the plots can not be trusted? We are guaranteed to find that the percent signal change will be greater than zero (since regions would be selected only if the average response to the three motion conditions was greater than fixation). For regions that were selected for responding more to face than radial motion, we are sure to find such an effect. For regions that were selected for responding more to hand rather than radial motion, we are sure to find that as well. We are also guaranteed to find that these differences will be greater than the noise fluctuations in the data – such is the definition of statistical significance. Furthermore, the magnitude of the guaranteed effects cannot be trusted, because the selection process is also likely to select voxels with random noise favoring these effects.

Nevertheless, this practice is very common – it is rare to find an issue of a neuroimaging journal in which none of the articles have plotted non-independent data. Indeed, even one of the authors of this chapter has done this previously (Culham et al., 1998; Kanwisher, McDermott, & Chun, 1997). This leaves us with several questions: What can plots of this sort contribute? Just how damaging is the practice of plotting the data that were used for selection?

2.2.3 What information may one glean from non-independent data?

In general, plotting non-independent data is misleading, because the selection criteria conflate any effects that may be present in the data from those effects that could be produced by selecting noise with particular characteristics. On the other hand, plots of non-independent data sometimes contain useful information orthogonal to the selection criteria. When data are selected for an interaction, non-independent plots of the data reveal which of many possible forms the interaction takes. In the case of selected main effects, readers may be able to compare the activations to baseline and assess selectivity. In either case, there may be valuable, independent and orthogonal information that could be gleaned from the time-courses. In short, there is often information lurking in graphs of non-independent data; however, it is usually *not* the information that the authors of such graphs draw readers' attention to. Thus, we are not arguing against displaying graphs that contain redundant (and perhaps biased) information, we are arguing against the implicit use of these graphs to convince readers by use of non-independent data.

2.2.4 How damaging are plots of non-independent data?

In cases where no inferential statistics are computed on the selected data[2], conclusions are explicitly based on the voxel selection process itself, and not the displayed ROI analyses. In such cases plotting these graphs is a problem only because they may mislead readers. The reader is presented with graphs that appear informative (in so far as they show data exhibiting effects that the paper is describing), but the graphs are not informative of the primary claims, and are distorted by selection bias. Authors that show such graphs must usually recognize that it would be inappropriate to draw explicit conclusions from statistical tests on these data (as these tests are less common), but the graphs are presented regardless. Unfortunately, the non-independence of these graphs is usually not explicitly noted, and often not noticed, so the reader is often not warned that the graphs should carry little inferential weight.

Just as in the case of testing for a selected effect, a particularly troublesome situation arises when voxels are selected at a low threshold – a threshold too low to effectively correct for multiple comparisons. In these cases, the displayed graphs falsely bolster the reader's confidence in the reliability of the effects, while in such cases, the reader's confidence in the result should be based only on the threshold used to initially select the voxels.

Because the most damaging consequence of plotting non-independent data is misled readers, a good antidote is full disclosure. Authors should explicitly state whether the plot corresponds to the same (non-independent) data used for the selection criteria, or different, independent data. Furthermore, if non-independent data are plotted, this should be accompanied by a description of which effects are biased and statistically expected due to the selection criteria. Ideally, any such graph would also feature an illustration of these biases (Baker, Simmons, Bellgowan, & Kriegeskorte, 2007).

[2] Although see page 970 of Thompson et al., there is an implication that statistical tests were run on the extracted signal change. The presumed inferential weight of the plots is further suggested by the presence of error bars.

2.3 Reporting correlations in voxels selected for correlations

A recent methodological trend, especially in social cognitive neuroscience, is the correlation of evoked activity with some traits of the individual. One such example is (Eisenberger, Lieberman, & Williams, 2003). In this study, participants played 'Cyberball', a video game in which subjects tossed a virtual ball with other agents in the game. Subjects were told that the agents were other human participants, while in reality they were simulated computer characters. On some runs subjects watched the other agents play the game, in another run subjects watched the other agents play the game while apparently intentionally excluding the subject. The researchers identified regions in the anterior cingulated cortex (ACC) that were more active during trials when subjects were excluded from the game than trials when they were included (excluded-included contrast). Within this contrast, the researchers then found brain regions where activity was correlated with ratings of distress elicited from each subject after the experiment. The authors report two sets of coordinates in the ACC that correspond to a positive correlation between BOLD and self-reported distress. They also report impressively high correlations at each of these coordinates: r=0.88 and 0.75.

What do these correlations mean? They are the correlations of voxels selected for having correlations significant at the $p<0.005$[3] level. The significance of a correlation may be assessed by a t-test with a t-value of r^2/(1-r^2)/N-2. Given that there were 13 subjects in the experiment, we can compute the minimum correlation necessary for voxels to be selected. This minimum possible r value is 0.7[4], so we know we will see an average r value greater than 0.7 in any voxels that were selected, even if the null-hypothesis holds true for those voxels (these would be voxels selected due to noise that aligned with the selection criteria).

Surely, you might suppose, since some number of voxels were above the critical correlation values necessary to reach significance, activity in the ACC must truly be correlated with self assessed distress. We do not aim to question this finding – if the assumptions of the minimum cluster size method for multiple-comparison correction were met in the multi-voxel analysis undertaken in this paper, there may indeed be a true correlation. We object, however, to the prominent display of average correlations from voxels that were selected for having significant correlations.

Imagine that ACC activity is correlated with subjective distress. This means that all voxels in the ACC (as identified by the excluded-included contrast) have some greater than zero correlation with subjective distress. The correlations in each of these voxels will be perturbed by noise: by chance, some voxels will cease to have detectable correlations, while other voxels, by chance, will become more correlated. All of the voxels in the ACC will follow some distribution of correlation values. An independent analysis of these correlations could have averaged the correlations across all voxels in the ACC, and computed statistics on this quantity. However, instead, the authors found regions within the ACC that were significantly correlate with subjective distress. Thus, from the distribution of all voxels and their respective correlations, the authors chose only those that had correlations greater than 0.7, then averaged them. Such a procedure is guaranteed to find an average correlation greater than 0.7, even if the true

[3] This was the threshold used in conjunction with a minimum cluster-size constraint. The assumptions and validity of this particular correction for multiple comparisons may be disputed, but here we are concerned with how the selection criteria this method imposes on the data affects the subsequent analyses.

[4] If one computes the inverse of the Fisher Z-transform method for ascertaining significance of a correlation, the numbers work out even less favorably.

correlation between evoked ACC activity and subjective distress is substantially lower. Again, if the multiple comparisons selection was done appropriately, it is still likely that the ACC does contain such a correlation; however, the magnitudes of the average correlations the authors report are spuriously elevated.

We have dwelt on this example because, unlike the post-hoc displays of signal change described previously, biased post-hoc displays of correlations seem to be substantially more convincing to audiences and readers, and appear to appeal to high profile journals (Dapretto et al., 2006; Kennedy, Redcay, & Courchesne, 2006; Mobbs, Hagan, Azim, Menon, & Reiss, 2005; Yoo, Hu, Gujar, Jolesz, & Walker, 2007). Since biased analyses and results such as these have a greater impact on audiences, it is more important to be aware of them, and to curb their use.

2.4 Multivariate correlations

Another newly popular class of analyses are even more conducive to hidden non-independence errors: multivariate analyses. In these analyses (e.g., (Haxby et al., 2001)), researchers assess the multivariate pattern of voxel activation for any given condition. That is, to what extent is the pattern of increased and decreased BOLD signal across voxels in a particular region (a measure independent of the mean signal in that region) diagnostic of a particular condition? In Haxby's analysis, this was measured as the correlation of voxel activations across two sets of identical conditions compared to two sets of different conditions. When correlations between identical conditions are greater than correlations between different conditions those conditions may be distinguished by the pattern. This intuition has been extended into more elaborate machine learning methods that explicitly classify conditions based on the evoked patterns of activation.

Just as in standard analyses, researchers typically select some subset of voxels on which to perform a multivariate analysis (to characterize a particular cortical region, gain power or remove uninformative voxels). Unfortunately, in the original Haxby paper, the method used to select voxels was not fully independent from the subsequent analysis. While this is not likely to have strongly affected the main results of that study, it is worth explaining the problem as an illustrative case.

Haxby et al. selected voxels based on significance in an omnibus ANOVA across all stimulus conditions, which was computed on all data (to be split into odd and even runs for the pattern analysis later). An omnibus ANOVA is significant insofar as one or more of the group means is different from the others. Effectively, this selection criterion biases the final correlations one might obtain: voxels will be selected if their mean activation is significantly different in one condition than another (and this would have to be reliable, across both datasets). If one condition is reliably different from others within this voxel, this means that activation across split halves will be better correlated for identical than different conditions.

Of course, the strength of this bias depends on how much the conditions differ from fixation. In the Haxby et al. paper, most of the reported effect is likely driven by true underlying effects. However, the fact that the analysis could be biased is a nontrivial problem that can produce spurious results (Simmons et al., 2006).

2.5 Summary

We have described four classes of analysis that are tainted by the non-independence error. In some of the case studies, the error undermined the main claims of the paper, in other cases, it simply resulted in the display of redundant information. Our goal in this section was not to single out these particular papers – many other examples are available. Our goal was to illustrate the many faces of the non-independence error in fMRI research. We hope that in describing these cases, we have provided a broad enough spectrum such that readers may be able to generalize to new instances, and spot these errors when planning experiments, writing papers, and reviewing for journals.

3. Why the non-independence error is prevalent in fMRI

The non-independence error we have outlined is not novel and has been committed in many other disciplines; however, it seems to be especially prevalent in fMRI. For example, five of the eight fMRI studies in a recent special issue on "Social Cognitive Neuroscience" included non-independent analyses (Neuroimage, 2005, 38; (den Ouden, Frith, Frith, & Blakemore, 2005; Gillath, Bunge, Shaver, Wendelken, & Mikulincer, 2005; Harris, Todorov, & Fiske, 2005; Mitchell, Banaji, & Macrae, 2005; Sander et al., 2005)). There are three circumstances of neuroimaging that put the field at high risk. First, fMRI researchers work with massively multidimensional datasets, in which only a subset of dimensions contain information that may be relevant to the experiment. This situation encourages researchers to select some subset of their data for analysis, thus to use non-independent selection criteria. Second, fMRI analyses are complicated, involving many steps and transformations before the final statistics may be computed, resulting in confusion (and thus a diminished ability to identify such errors) not only on the part of the researchers themselves, but also on the part of reviewers. Finally, fMRI research usually asks binary qualitative, not quantitative, questions – data are presented as binary values (significant or not significant) further veiling any biases that may lie behind the analysis.

3.1 fMRI data are massively multidimensional.

A typical 'low resolution' scan on a low-field magnet will produce an imaging volume every 3 seconds. The imaging volume will contain 20 3mm slices, each of which is divided into a 64x64 (3mm x 3mm) grid, producing 81,920 measurements every 3 seconds. A 'high resolution' scan on state-of-the-art scanners might produce an image volume every 2 seconds, and this volume may contain 30 1.5mm slices, each of which is divided into a 128x128 (1mm x 1mm) grid, producing a staggering 491,520 measurements every 2 seconds. Thus, a single scan session could easily produce more than 1 billion measurements, and often multiple sessions are combined in the analysis.

Statisticians are not known to complain about an overabundance of data, and the problem here is not the raw number of measurements, but rather the fact that usually only a small proportion of the measurements are informative about the experimental question. In a fortuitous and skillfully executed experiment, one may find 5% of voxels to be of experimental interest. This poses a difficult multiple comparisons problem for 'whole-brain' analyses. In this chapter, we have only indirectly discussed this problem, because the applications (and misapplications) of the many technical methods used to correct for multiple-comparisons are a considerable topic on their own. Instead, we have discussed a consequence of this problem: selection.

When experimenters ask subtler questions than 'which area lights up under condition X?', they invariably select some subset of the enormous fMRI dataset to avoid correcting for multiple comparisons and losing statistical power. Therefore, most modern fMRI analyses proceed in two stages: (1) identifying a subset of voxels that play an interesting role in the experiment (a region of interest -- ROI)[5], then (2) assessing some additional measure in those voxels. Obviously, the criteria used for selection in step 1 are a condition one puts on the measure in step 2 – in this chapter, we have discussed whether the conditions from step 1 satisfy the assumption of independence necessary for the statistical analyses in step 2.

The non-independence error arises from the relationship between the ROI selection method and the statistical test. If the conditions imposed by the selection process alter the distribution assumed by the null hypothesis of the subsequent statistical test, then this secondary test is non-independent. Naturally, this will mean that some combinations of ROI selection methods and analyses do satisfy the independence assumption (and are hence legitimate), and different combinations of the same techniques may not (and are not).

Selecting small subsets of large datasets is an integral part of most fMRI analyses to a much greater degree than in behavioral studies. Since 'selection' (biased or not) is more common in fMRI, then, even if selection biases are inadvertently introduced equally often in analyses in other fields, we would expect to see a greater proportion of reported results tinged by selection bias in fMRI.

3.2 fMRI analyses are complicated (both to do and to review).

There are many steps between the acquisition of fMRI data and the reported results. Before the final analysis, a staggering variety of pre-processing techniques are applied to the data. The four-dimensional image (volume by time) obtained from the scanner may be motion-corrected, co-registered, transformed to match a prototypical brain, resampled, detrended, normalized, smoothed, trimmed (temporally or spatially), or any subset of these, with only a few constraints on the order in which these are done. Furthermore, each of these steps can be done in a number of different ways, each with many free parameters that experimenters set, often arbitrarily. The decisions an experimenter makes about preprocessing are less likely to be crucial for the issue of non-independence[6]. However, these steps play an important role in the final results, and must be specified when describing an experiment.

After pre-processing, the main analysis begins. In a standard analysis sequence, experimenters define temporal regressors based on one or more aspects of the experiment sequence, choose a hemodynamic response function, and compute the regression parameters that connect the BOLD signal to these regressors in each voxel. This is a whole brain analysis (step 1 described in section 3.1), and it is usually subjected to one of a number of methods to correct for multiple comparisons (False detection rates, minimum cluster size thresholds, Bonferroni, etc.). Because it is difficult to gain enough power for a fully corrected whole brain analysis, such analyses are rarely done in isolation. Instead, in conjunction with anatomical assumptions, the

[5] Note that defining a region of interest need not be done with a priori functional localizers (for a discussion of this controversy, see (Friston, Rotshtein, Geng, Sterzer, & Henson, 2006; Saxe, Brett, & Kanwisher, 2006)) – this may be done with orthogonal contrasts from the present experimental manipulation, or even anatomy.

[6] However, an often ignored fact is the key role played by voxel size and smoothing parameters in the assumptions behind minimum cluster-size methods for multiple-comparisons correction – thus, smoothing, at least, alters the conditions imposed by the ROI selection analysis.

whole brain analysis is often the first step defining a region of interest in which more fine-grained, technically sophisticated, and interesting analyses may be carried out (step 2 in section 3.1).

The analyses within selected ROIs may include exploration of timecourses, voxel-wise correlations, classification using support vector machines or other machine learning methods, across-subject correlations, etc. Any one of these analyses requires crucial decisions that determine the soundness of the conclusions. Importantly, it is the interaction between a few of these decisions that determines whether or not a statistical analysis is tarnished by non-independent selection criteria.

The complexity of the fMRI analysis has two consequences, each of which can only increase the likelihood that experimenters will inadvertently use non-independent selection criteria. First, with so many potential variables, it is difficult to keep track of possible interactions that could compromise independence. Second, and perhaps more important, to fully specify the analysis in a publication requires quite a lot of text – text that high profile journals prefer not to use on Methods sections. So editors (and publication policies) encourage authors to exclude details of the analysis on the assumption that they may be trivial or unimportant. The result is a hamstrung review process in which reviewers are not given the full information necessary to evaluate an analysis. The complexity of fMRI analyses is not inherently bad; however, the complexity offers opportunities for researchers to make mistakes and diminishes opportunities for reviewers to spot the errors.

3.3 fMRI analyses are usually qualitative

The qualitative nature of the questions asked and results obtained in fMRI also contributes to the prevalence of the non-independence error. An area is said to respond differently, or not; to contain some information, or not; to predict behavior, or not. Of course, the brain states underlying the effects observed are quantitatively different, and we draw arbitrary lines to produce qualitative answers. Why does this matter?

As our examples have shown, the non-independence error in fMRI analyses usually does not guarantee a particular result. Instead, the results are biased to favor a particular outcome. The extent to which results are biased is usually unclear. Since results are displayed as binary outcomes (significant or not), it is substantially more difficult to evaluate whether the significance of an effect is due to the bias. One might ask what proportion of an effect is suspect, but such a question arises less naturally for results with binary outcomes. By drawing hard thresholds, the practice of significance testing further muddies the results of an analysis, and complicates evaluation.

3.4 Summary

fMRI is not the only field to contain biased results and non-independent selection criteria, and it is also not the only field to suffer from the conditions previously described. Gene-sequencing involves massively multidimensional data. Electrophysiology experiments require complicated time-series analysis. Most behavioral experiments in psychology evaluate results via a statistical test with a binary outcome. Although these factors are shared by other fields (and result in non-independence errors in those fields), fMRI data and analyses are subject to all of these factors, thus increasing the odds that any one analysis may be tainted.

4. Heuristics for avoiding non-independence errors

How might one avoid committing variants of the non-independence error when conducting fMRI analyses? For independence of selection and analysis, we require that selection criteria, if applied to a distribution from the null hypothesis, will produce another distribution drawn from the null hypothesis. Three classes of solutions seem intuitively reasonable. The best option is to safeguard against non-independence by using different datasets for the selection and analysis. Another possibility is to determine *a priori* whether the analysis and selection criteria are independent. Rather than deducing this independence analytically, a third option is to assess such independence by simulation

Each of these strategies has advantages and disadvantages and none can be guaranteed to be effective. Although we advocate the use of independent data, it is important to note that even then, some degree of the other two approaches may be required of the researcher.

4.1 Ensuring independence by separating datasets.

Perhaps the most intuitive precaution against non-independent selection criteria and analysis is the use of completely different data sets. Voxel selection would be based on a subset of the data (specific trials, blocks, runs, experiment halves, etc.) while the analysis would be performed on the remaining data. If the data are truly independent, selection bias cannot be introduced when selection and analyses are executed on different subsets of the data.

However, certain divisions of the data and physiological factors may render superficially independent data actually non-independent. Imagine that we decide to separate our data into odd- and even-numbered columns of the fMRI image. We will select a subset of even numbered columns for further analysis based on what would be otherwise a non-independent criterion imposed on paired odd-numbered voxels. The data-set used to define a region of interest and that used for the subsequent analysis are *nominally* independent. However, in this case the data are not really independent due to the spatial correlation intrinsic to fMRI (as should be expected from either explicit spatial smoothing, or the correlation induced by vasculature, blood flow, and MR image construction).

One could imagine a different example, in which alternating image acquisitions (TRs) are used to select voxels, and the interceding images are used for the subsequent analysis. Explicit temporal smoothing of the data is quite rare in fMRI, so non-independence is not likely to be introduced from pre-processing. However, again physiology introduces bias: due to the temporal delay and extent of the hemodynamic response function, temporally contiguous images are far from independent.[7]

These two examples demonstrate that the use of distinct datasets for selection and test does not guarantee independence. There are myriad ways in which data may be rendered more or less independent by preprocessing, non-random experimental sequence, etc.

4.2 Evaluating independence by analytical deduction.

[7] See (Carlson, Schrater, & He, 2003) for an example of such an analysis, as well as some discussion about the magnitude of the presumed bias.

One might attempt to deduce, a priori, whether the conditions one imposes on the sample of voxels selected for further analysis are orthogonal to the analysis itself. Since analytical solutions to this problem will often be intractable, and assumptions about the joint probability distribution of all of the data will be unjustified, we do not consider the possibility of attempting to derive independence via pure mathematics. That said, the only method we know for determining independence a priori is to try very hard to find reasons why the selection criteria might not meet this criterion, and fail to find any. It seems perilous to advocate such a subjective use of statistics (after all, some people may not find failures of orthogonality where others succeed). Indeed, the cases we have described, more likely than not, reflect a failure to come up with a reason why the orthogonality condition is not met.

4.3 Assessing independence by numerical simulation.

Rather than producing arm-chair arguments about the independence of selection from the subsequent analysis, one may run numerical simulations to measure mutual information between the conditions and the null hypothesis distribution (described in section 1.4). This approach occurs in the literature most frequently as a post hoc illustration of a failure to meet the non-independence criterion (Baker, Hutchison, & Kanwisher, 2007; Simmons et al., 2006). Such post hoc refutations of biased analyses are useful in weeding out spurious results from the literature and advancing science. However, we hope that authors will take it upon themselves to use such approaches to determine the soundness of their own analyses before they are published. Permutation tests are one particularly effective method when analyses are particularly complicated. Researchers can randomly permute the condition labels for their data and undertake the same analysis. If this is done enough times, it is possible to empirically estimate the probability of the outcome observed with the true data labels. Unlike simpler (and faster) white-noise simulations, this permutation analysis includes the non-gaussian structure of the BOLD signal noise, and is thus more accurate.

4.4 Summary

All in all, we would advocate using one dataset for voxel selection and a different, independent dataset for subsequent analysis to decrease the likelihood of the non-independence error. However, due to spatiotemporal correlations in fMRI data, even in these cases, independence is not guaranteed, and researchers ought to use caution to make sure the two datasets are in fact independent.

It is worth noting that we are explicitly advocating the use of independent data, not necessarily alternative stimulus sequences for use in localizers. This advice is orthogonal to the "ROI debate" (Friston, Rotshtein, Geng, Sterzer, & Henson, 2006; Saxe, Brett, & Kanwisher, 2006) about the role, and meaning, of functionally defined regions. However, we do depart from Friston et al. in advising that independent data be used. Friston advocated the use of a 'factorial design' such that voxel selection and a subsequent analysis are achieved with the same dataset (with the condition that voxel selection and the analysis are orthogonal). In principle, if these analyses are truly orthogonal then they are independent, and we agree. Unfortunately, orthogonality of selection methods and analyses seems to be often wrongly assumed[8]. While it is

[8] Consider the case of selecting voxels based on two main effects and testing for the interaction. Although the two main effects and the interaction appear orthogonal, they are not. If we select random noise in which two main

more economical to use one dataset for selection and analysis, it seems much safer to use independent datasets (indeed, if spurious results are published due to failures of orthogonality, the entire research enterprise ends up substantially more costly).

5. Closing

We have described a common error in fMRI research: the use of non-independent selection criteria and statistical analyses. This error takes many forms, from seemingly innocuous graphs that merely illustrate the selection criteria rather than contribute additional information, to serious errors where significant results may be produced from pure noise. In its many forms, this error is undermining cognitive neuroscience. Public broadcast of tainted experiments jeopardizes the reputation of cognitive neuroscience. Acceptance of spurious results wastes researchers' time and government funds while people chase unsubstantiated claims. Publication of faulty methods spreads the error to new scientists. We hope that this chapter finds itself in the hands of the authors, reviewers, editors, and readers of cognitive neuroscience research and arms them with the formalism and intuition necessary to curtail the use of invalid, non-independent analyses.

---

effects are significant ([A1+A2] > [B1+B2], and [A1+B1] > [A2+B2]), the mutual constraint of *both* main effects will preferentially select positive noise in the A1 cell and negative noise in the B2 cell, thus biasing results toward an interaction.

Appendix: Formal description of the non-independence error

What exactly is the error that leads to the absurd conclusion in the card example from the introduction? Here we describe it in three additional theoretical frameworks: propositional logic, probability theory, and information theory.

1.1 Propositional Logic

In formal logic, the non-independence error goes under many names: *petitio principii*, "begging the question", or "circular reasoning"[9]. A distilled example of begging the question will read as follows:

p implies p;
suppose p;
therefore p.

In practice, of course, the fallacy is usually cloaked by many logical steps and obfuscatory wording, such that the assumption of p is rarely obvious. In the card (or evolution) example we start with the outcome (the particular arrangement of cards or genes: "suppose p"), and then marvel that we have found the same outcome ("therefore p"). Thus, these cases exemplify 'begging the question': a condition when the conclusion is implicitly or explicitly assumed in one of the premises.

In the introduction, and throughout the paper, we concern ourselves with statistics and probability, so this fallacy is best fleshed out in terms of probability theory and statistics. However, the essence of the problem is still simple question begging: evaluating the truth of a statement which has been presupposed.

1.2 Probability Theory

We can also analyze begging the question in a probabilistic framework by evaluating the implicit and explicit assumptions in the deck of cards example. It is true that a fully shuffled deck of cards has one of 52 factorial possible arrangements. It is also true that a deck of cards randomly sampled from a distribution over all possible shufflings will be unlikely to have any one particular arrangement (a probability of 1/52 factorial). So if we were to choose a random arrangement and then shuffle a deck of cards until we found that arrangement, we would probably find ourselves busy for a long time.

However, we are not evaluating the prior probability of a random deck of cards having a random arrangement. Instead, we are evaluating the probability that a deck of cards will have a specific order: the order we just observed the deck of cards to have. Thus, the probability distribution we should be evaluating is not the prior probability $P(X=x)$, but the conditional probability $P(X=x \mid X=x)$. Of course, this probability is 1.

One of an enormous number of outcomes was possible, but if we condition on the observed outcome, that particular outcome is guaranteed. In the formalism of probability this difference between prior and conditional probabilities may be described as a violated assumption of independence: $P(X)$ is not equal to $P(X|C)$, where C is our condition.

The deck of cards case is an extreme example where the disparity between the assumed prior probability and the relevant conditional probability is particularly large – the prior

[9] In our discussion, we consider these three interchangeable, but some differentiate "begging the question" as an error that occurs within one argument, while "circular reasoning" involves two mutually dependent arguments.

probability is impossibly low and conditional probability is 1. These probabilities make the scenario easy to describe in terms of predicate logic, but the same violation of independence arises if the probabilities are not 1 and (near) 0.

1.3 Information Theory

Finally, we can formalize the non-independence error in the framework of information theory to appeal to another intuition about a desirable quality of statistical analyses: how much information they assume of, or add to, the data. If we specify the null hypothesis of our statistical test as P(X) and the selection criteria as imposing a condition on the data, producing P(X|C), then we can derive how much information the selection criteria (condition C) give us about X.

The Shannon entropy of a random variable reflects the uncertainty present in the probability distribution of that random variable.

$$H(X) = -\sum P(X)\log_2 P(X)$$

Intuitively, this number expresses how many bits would be necessary to encode a sample from that distribution, thus expressing uncertainty in terms of information.

With this measure we can describe how much information (H) a selection condition gives us about a random variable by evaluating how much less uncertainty there is in the conditional distribution compared to the prior distribution. This is expressed as mutual information: I(X;C) = H(X) – H(X|C).

To return to our example of a deck of cards, we can assess how many bits of information it takes to encode the order of a random deck of cards:

$$H(X) = -\sum_{i=1}^{52!} \frac{1}{52!}\log_2 \frac{1}{52!} = -\log_2 \frac{1}{52!} = 226$$

And we can calculate the amount of information necessary to encode a deck of cards sampled from the distribution of all decks of cards which we have observed to have a particular order (index 1):

$$H(X \mid C) = -1\log_2 1 - \sum_{i=2}^{52!} 0\log_2 0 = 0$$

This means that given the selection criterion, we get no additional information.

The mutual information is thus 226 bits (226-0). This is an enormous number, reflecting the huge disparity between P(X) and P(X|C). It is also useful to express the information gained from our selection criteria as a proportion of all the information one could have gained:

I(X;C)/H(X) = 226/226=1.

In this extreme example, our selection criteria have given us all the information available about a sample from the prior distribution. Our sampling process has thus fully constrained our data by giving us full information about it. Obtaining full information about the outcome from the starting conditions is identical to begging the question in propositional logic: starting with full information about the outcome means that the outcome was presupposed.

1.5 Summary

We formalized the non-independence error from the introduction in terms of propositional logic, probability theory, statistics (in Section 1 of the main text), and information

theory. This allowed us to describe violations of assumed independence in probability theory as a generalized case of 'begging the question' in propositional logic. Similarly, the error of 'selection bias' in classical statistics is formally equivalent to a violation of independence in probability theory. We then used information theory to quantify violations of independence as the mutual information between the measurement and the selection criterion. Finally, by taking the limiting case wherein the mutual information between the measure and the selection criterion is 100% of the total information available in the measure, we again produce the case of explicitly begging the question in propositional logic. By describing this error in different frameworks we hope that readers can apply their intuitions from any of these domains to actual fMRI examples.

We use the term 'non-independence error' throughout this paper in favor of 'begging the question' to convey the idea that the selection criteria need not be so restrictive as to guarantee the outcome (as is the case in propositional logic). Instead, if the selection criteria applied to a null hypothesis distribution alter the distribution in any way, they are introducing some degree of sampling bias, providing some amount of information about the outcome, and thus will produce biased results due to the violation of independence.

Baker, C. I., Hutchison, T. L., & Kanwisher, N. (2007). Does the fusiform face area contain subregions highly selective for nonfaces? *Nat Neurosci, 10*(1), 3-4.

Baker, C. I., Simmons, W. K., Bellgowan, P. S., & Kriegeskorte, N. (2007). *Circular inference in neuroscience: The dangers of double dipping.* Paper presented at the Society for Neuroscience, San Diego.

Cantlon, J. F., Brannon, E. M., Carter, E. J., & Pelphrey, K. A. (2006). Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biol, 4*(5), e125.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *J Cogn Neurosci, 15*(5), 704-717.

Culham, J. C. (2006). Functional Neuroimaging: Experimental Design and Analysis. In R. Cabeza & A. Kingstone (Eds.), *Handbook of Functional Neuroimaging of Cognition (2nd ed.)*. Cambridge, MA: MIT Press.

Culham, J. C., Brandt, S. A., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *J Neurophysiol, 80*(5), 2657-2670.

Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., et al. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nat Neurosci, 9*(1), 28-30.

den Ouden, H. E., Frith, U., Frith, C., & Blakemore, S. J. (2005). Thinking about intentions. *Neuroimage, 28*(4), 787-796.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science, 302*(5643), 290-292.

Gillath, O., Bunge, S. A., Shaver, P. R., Wendelken, C., & Mikulincer, M. (2005). Attachment-style differences in the ability to suppress negative thoughts: exploring the neural correlates. *Neuroimage, 28*(4), 835-847.

Grill-Spector, K., Sayres, R., & Ress, D. (2006). High-resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nat Neurosci, 9*(9), 1177-1185.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: neuro-imaging dispositional inferences, beyond theory of mind. *Neuroimage, 28*(4), 763-769.

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

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci, 17*(11), 4302-4311.

Kennedy, D. P., Redcay, E., & Courchesne, E. (2006). Failing to deactivate: resting functional abnormalities in autism. *Proc Natl Acad Sci U S A, 103*(21), 8275-8280.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *Neuroimage, 28*(4), 757-762.

Mobbs, D., Hagan, C. C., Azim, E., Menon, V., & Reiss, A. L. (2005). Personality predicts activity in reward and emotional regions associated with humor. *Proc Natl Acad Sci U S A, 102*(45), 16502-16506.

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53*(2), 293-305.

Rees, G., Frith, C. D., & Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. *Science, 278*(5343), 1616-1619.

Ruff, C. C., Blankenburg, F., Bjoertomt, O., Bestmann, S., Freeman, E., Haynes, J. D., et al. (2006). Concurrent TMS-fMRI and psychophysics reveal frontal influences on human retinotopic visual cortex. *Curr Biol, 16*(15), 1479-1488.

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., et al. (2005). Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. *Neuroimage, 28*(4), 848-858.

Simmons, W. K., Matlis, S., Bellgowan, P. S., Bodurka, J., Barsalou, L. W., & Martin, A. (2006). Imaging the context-sensitivity of ventral temporal category representations using high-resolution fMRI. *Society for Neuroscience Abstracts*.

Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., & Hirsch, J. (2006). Predictive codes for forthcoming perception in the frontal cortex. *Science, 314*(5803), 1311-1314.

Thompson, J. C., Hardee, J. E., Panayiotou, A., Crewther, D., & Puce, A. (2007). Common and distinct brain activation to viewing dynamic sequences of face and hand movements. *Neuroimage, 37*(3), 966-973.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*(6984), 751-754.

Yoo, S. S., Hu, P. T., Gujar, N., Jolesz, F. A., & Walker, M. P. (2007). A deficit in the ability to form new human memories without sleep. *Nat Neurosci, 10*(3), 385-392.