Articles

Ethical and Legal Implications of the Methodological Crisis in Neuroimaging

PHILIPP KELLMEYER

Abstract: Currently, many scientific fields such as psychology or biomedicine face a methodological crisis concerning the reproducibility, replicability, and validity of their research. In neuroimaging, similar methodological concerns have taken hold of the field, and researchers are working frantically toward finding solutions for the methodological problems specific to neuroimaging. This article examines some ethical and legal implications of this methodological crisis in neuroimaging. With respect to ethical challenges, the article discusses the impact of flawed methods in neuroimaging research in cognitive and clinical neuroscience, particularly with respect to faulty brain-based models of human cognition, behavior, and personality. Specifically examined is whether such faulty models, when they are applied to neurological or psychiatric diseases, could put patients at risk, and whether this places special obligations on researchers using neuroimaging. In the legal domain, the actual use of neuroimaging as evidence in United States courtrooms is surveyed, followed by an examination of ways that the methodological problems may create challenges for the criminal justice system. Finally, the article reviews and promotes some promising ideas and initiatives from within the neuroimaging community for addressing the methodological problems.

Keywords: neuroimaging; reproducibility; replication; neuroethics; neurolaw

Introduction

In the past two decades, neuroimaging has become the dominant research method for investigating structure–function relationships of the human brain *in vivo*. Trawling any database of scholarly article reveals that, in this period alone, hundreds of thousands of studies using functional magnetic resonance imaging (fMRI) were published in peer-reviewed academic journals. Recently, however, the methods and software used for analyzing fMRI data have come under increased scrutiny. These investigations have, among other concerns, questioned the validity and reproducibility of statistical inference,[1,2] cross-software comparability,[3,4] and interpretation,[5] and have revealed the influence of image preprocessing methods on the anatomical assignment of brain activity,[6] all of which have an important impact on the functional interpretation of neuroimaging data.

The aim of this article is to explore some ethical and legal implications of this "methodological crisis" in neuroimaging that have hitherto remained underexplored. First, I will contextualize the problem by highlighting the broader "replication crisis" in experimental psychology, biomedicine, and other fields and by exploring antecedents from historical brain-based research methods that have relied on correlational methods. Second, I will discuss how the methodological problems may affect and compromise basic research in cognitive and clinical neuroscience research, particularly in vulnerable patient populations in neurology and psychiatry.

This work was supported by the German Ministry of Education and Research (BMBF) grant 13GW0053D (MOTOR-BIC) and the DFG grant EXC 1086 BrainLink-BrainTools to the University of Freiburg, Germany.

Cambridge Quarterly of Healthcare Ethics (2017), **26**, 530–554.
© Cambridge University Press 2017.
doi:10.1017/S096318011700007X

Third, I will consider how these problems could also spill over into the application of neuroimaging in the clinical routine.

In research areas in social psychology that increasingly use neuroimaging, the unreliability of fMRI-based analysis may lead to faulty models of human cognition, behavior, and personality. I outline the potential negative consequences of applying such models in a legal context, specifically for using neuroimaging-as-evidence in the courtrooms.

I conclude with a discussion of potential solutions for mitigating these ethical and legal consequences of the methodological problems in neuroimaging research. Specifically, I will: (1) promote ideas from within the neuroimaging community on how to improve the transparency and reproducibility in neuroimaging methodology; (2) highlight the need for researchers and science journalists to communicate accurately the uncertainty attached to neuroimaging research, and (3) highlight the need to increase funding of research on the ethical, legal, and social consequences of methodological problems in neuroimaging.

Low Levels of Confidence: The Current Replication Crisis in Science

Major Issues in the Replication Crisis in Science

In recent years, the quality of scientific research has come under intense scrutiny. From experimental psychology to neuroscience and biomedical research, many established research practices are called increasingly into question. These systematic efforts in assessing and comparing the quality of scientific research have eventually spawned an emerging research discipline of its own called "metascience."[7] Major issues in this discussion are: (1) questions on the replication of studies and effect sizes; and (2) the ecological validity of laboratory-based research; as well as the influence of (3) mostly unintentional systematic biases (e.g., publication bias); and (4) overtly unethical research practices (such as "p-hacking"[8] or "HARKing" [hypothesizing after the results are known][9]) on the overall quality and trajectory of empirical research. This "replication crisis," the shorthand for these problems, has now also reached the neuroimaging community in experimental psychology, cognitive neuroscience, and medicine. Before turning to the specific methodological problems in neuroimaging, I will first contextualize this crisis of confidence within the recent history of brain-based research.

Brain-Based Research: Victim of Its Own Success and the Perils of "Brain Overclaim Syndrome"

Most of the brain-based research on human cognition and emotion in the past 100 years has been produced in three—clearly intertwined and overlapping—disciplines: experimental psychology, cognitive neuroscience, and clinical neuroscience (neurology and psychiatry). These disciplines have arguably produced great insights into structure–function relationships in the brain: the networks for memory formation,[10] and the role of the amygdala for threat detection in the "fear network,"[11] to name just a few examples. In lockstep with these spectacular advances, the popularity of brain-based explanations for human behavior, dispositions, and personality (as opposed to social dynamics or environmental factors) has soared. One need only to consider the rise and subsequent foundering of

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

psychoanalysis as the dominant mode for interpreting human behavior to recognize this historical dynamic. The "Decade of the Brain", as the 1990s were designated by the United States Congress and United States President George H.W. Bush, is now firmly into its 27th year and still alive and well; or is it?

With the advent of functional neuroimaging, first positron emission tomography (PET) and later functional magnetic resonance imaging (fMRI) in the 1990s, this already-asserted dominance of brain-based research over behavioral psychology went from steady cruising speed into overdrive. In 2015 alone, 29,295 articles using fMRI were published in peer-reviewed journals.[12] Imaging has become an integral and crucial part of clinical assessment in neurology and psychiatry. The "success" of neuroimaging as a technology for research and clinical applications in the modern era might only be comparable to the rise of analytic methods in molecular biology, such as the polymerase chain reaction (PCR), or the ubiquity of the computer as a research tool in science.

This success of brain-based research, fueled by the persuasive and iconic power of "neuro images"—henceforth neuroimaging iconolatry, or neuroimagery for short—has precipitated a backlash from traditionalists of the behavioral and/or psychoanalytic persuasion as well as skeptics in other fields. This skepticism toward the uncritical reverence of neuroimagery has been described, in the context of using neuroscience research as evidence in the courtroom, rather succinctly as "brain overclaim syndrome."[13] As a researcher and clinician who is familiar with both the strengths and weaknesses of fMRI and other neuroimaging methods, I advocate for taking an intermediate position in which we should acknowledge and critically discuss the methodological limits and pitfalls of neuroimaging while also recognizing its strengths and benefits. First, I will briefly discuss the ways in which the replication crisis is affecting experimental psychology, cognitive neuroscience, and biomedical research as of late.

The Replication Crisis in Experimental Psychology, Cognitive Neuroscience, and Biomedical Research

Psychological science has a long history of unmasking, facing, and—in most cases—improving on significant setbacks and methodological crises in its history as a field. As reproducibility and replicability are such important pillars of the scientific method, psychology was (and still is) shaken to the core by subsequent earthquakes of failed study replications. Yes, many researchers may have voiced concerns over the reproducibility of experiments in psychology in the past; however, what really substantiated the debate in recent years was the publication of a large replication study by the Open Science Collaboration in August 2015.[14] In this study, the participating researchers repeated 100 studies that were published in high-ranking journals in the year 2008 with the aim of replicating the results. The outcome of this large-scale collaborative effort was, to say the least, unexpected if not chastening. When combining the results from the original studies and their respective replications, the percentage of studies with a significant result dropped from 97 to 68 percent, and the mean effect size in the replication studies was only half the effect size of the original studies.

To take another recent example, together with other researchers, metascience pioneer John P.A. Ioannidis of Stanford University has recently published a study in which the authors looked at the distribution of effect sizes and estimated power

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

in 3,801 recent studies in cognitive neuroscience and psychology.[15] They found small median effect sizes ($d = 0.93$), and that studies in cognitive neuroscience had, on average, even lower statistical power than those in psychology. Importantly, they also found that slightly more than 50 percent of findings that were reported to be significant in these studies were likely to be false positives, replicating similar previous findings from metascientific studies in biomedical research.[16]

I should note, nevertheless, that some researchers have questioned whether there really is a serious replication crisis at all. Most of these counterarguments express doubts about the ideal of replication itself as a useful standard of scientific quality. Some researchers have claimed that the perceived crisis is based on an "epistemological misunderstanding"[17] between replicating a phenomenon or experimental effect (e.g., a particular cognitive bias) and particular mechanisms (the ways in which this bias may be produced). Others have argued that in the age of Big Data with complex and very large data sets that are often analyzed with adaptive algorithms (e.g., based on machine learning), the standard of reproducibility should really only apply to the method (e.g., publishing the code with which the data was analyzed) and not the exact results across studies.[18] As this debate is ongoing and remains, for the time being, unresolved, the examination here is predicated on the conviction that there is indeed a profound and relevant crisis in research methodology.

Therefore, I will now examine which aspects of this wide-ranging replication crisis in science spill over into neuroimaging research, and which methodological problems derive from neuroimaging itself.

The Methodological Crisis in Neuroimaging

Before I scrutinize the methodological problems specific to neuroimaging, I will define which techniques I refer to here as "neuroimaging." This is important, in my view, to delineate clearly the scope and limits of subsequent considerations. Second, I feel obliged to make a few remarks on some inherent epistemological problems that are common to all (brain) research that is based on correlational methods, as this has some bearing on how heavily (or lightly) we judge the impact of the methodological crisis.

Defining Neuroimaging and its Applications

I may begin by asking which recording techniques and analytical methods are referred to as neuroimaging and why? This is by no means a trivial exercise, as there is no clear textbook definition of which techniques qualify as neuroimaging and which do not.

Today there are many ways for recording biological signals from the brain. Researchers and clinicians use computed axial tomography/computed tomography (CAT/CT) scans to obtain x-ray-based (and sometimes contrast-enhanced) structural images of the brain. Structural MRI (sMRI) is used for acquiring high-resolution structural images and fMRI (or PET) is used for indirectly measuring task-related or resting-state brain activity. Researchers can also record electric field activity of populations of neurons from the scalp with classic electroencephalography (EEG) or from the brain surface with electrocorticography (ECoG), measure the minute magnetic fields induced by this electric activity with

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

magnetoencephalography (MEG), or measure the local blood flow optically (with optrodes) from the scalp with near-infrared spectroscopy (NIRS). These recordings (and subsequent analyses) can be superimposed on structural images of a subject's brain. Although these data from electrophysiological or hemodynamic measurements have different temporal and spatial characteristics (from each other and from fMRI or PET), each method provides valuable information for observing brain activity, whether in a relaxed "resting state" or in relation to specific tasks. Within this ecosystem of modern brain-research methods, I propose that one may distinguish a *narrow* and a *broad* view of neuroimaging.

The *narrow* view encompasses all techniques in which actual images of the brain are acquired: structural imaging for voxel-based symptom-lesion mapping (VBSLM) or voxel-based morphometry (VBM), and fMRI, diffusion tensor imaging (DTI), and PET.[19] The *broad* view also includes all methods in which electrophysiological (EEG, ECoG, MEG) or hemodynamic (functional NIRS [fNIRS]) signals are recorded, which are then superimposed on structural brain images. As fMRI also relies on the measurement of a hemodynamic brain signal, one might ask whether it belongs into the narrow or broad category. The salient difference here, I submit, is that in fMRI, the blood-oxygen level dependent (BOLD) signal is measured from acquired brain images, so-called "echo-planar images" (EPIs), whereas in fNRIS, EEG, ECoG, or MEG, the structural images on which the signals are later superimposed are acquired on a separate occasion. Therefore, for the benefit of limiting this discussion, I will mainly discuss the different radiographic techniques for obtaining structural and/or functional images of the brain—the narrow view of neuroimaging—as summarized in Table 1.

Sources of Variance and Error in Neuroimaging Research

As in all research fields, it is important to consider and acknowledge the sources of variance that contribute to the varying overall research quality in neuroimaging at different levels. At the measurement level, fluctuations in the quality of the data themselves (e.g. random changes in a subject's brain perfusion unrelated to neural activity) and the data acquisition process (e.g., inhomogeneous magnetic fields in the scanner) can all affect the quality of the acquired data. Many of the basic machine-level sources of variance contribute random fluctuations in signal quality, increasing the background noise and making it therefore more difficult for researchers to extract statistically sound and replicable results from the data. These problems are usually quite difficult to solve by any individual researcher in an experiment, but may be addressed at the engineering level of designing and building the measurement equipment, for example an MRI scanner.

What about sources of variance that *can* be influenced by the neuroimaging researcher? Variance also occurs at the level of experimental design; for example, the nature of the task and sequence of stimuli that a researcher chose for a particular study. Consider also parameters such as the sample size used in an fMRI study, the selection process of participating subjects, or the sex balance. Data analysis is another source of researcher-related variance and error; for example, when researchers are poorly trained in using the right software and settings.

Here I argue—following previous suggestions from colleagues—that the level of experimental design and data analysis is the most accessible and feasible level for improving the quality of neuroimaging research collectively.[20] Before looking

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

Table 1. List of Neuroimaging Methods

Method	Scanner	Key principle	Main uses
CAT/CT	CAT/CT	Combining axial radiographic images of the brain	Research: Because of radiation exposure, only used in clinical research Clinical: Widely used in neurology (and psychiatry) for stroke and many other scenarios Legal: Used to investigate anatomical brain abnormalities
fMRI	MRI	Measuring blood-oxygen level dependent (BOLD) signal from voxels in the brain to measure ongoing brain activity	Research: Cognitive and clinical neuroscience Clinical: Mapping of functionally salient brain regions before brain surgery Legal: Rarely used for investigating brain dysfunction or for brain-based lie detection
sMRI	MRI	T1-weighted imaging of the brain for obtaining stationary structural images in high quality	Research: Cognitive and clinical neuroscience Clinical: Widely used for imaging brain-based disorders (stroke, neurodegenerative disorders, depression) Legal: To investigate anatomical brain abnormality
DW-MRI	MRI	Measuring the diffusion of water molecules in the brain to image white matter fiber tracts	Research: Cognitive and clinical neuroscience Clinical: Mapping of important white matter fiber tracts before brain surgery Legal: Very rarely used to investigate for structural white matter abnormalities
VBM/ SBM/ VBSLM	MRI (CAT/ CT for VBSLM)	CAT scan or T1-weighted sMRI image is correlated with a behavioral measure (e.g., verbal fluency)	Research: Cognitive and clinical neuroscience Clinical: Sometimes used for differentiating patterns of atrophy in neurodegenerative diseases Legal: VBM in some cases to demonstrate brain abnormalities
PET	PET	Detection of gamma rays from a radionuclide in the bloodstream to image local cerebral metabolism	Research: Cognitive and clinical neuroscience Clinical: Imaging of brain tumors, neurodegenerative diseases (e.g. AD, PD, and others) Legal: Demonstrating brain dys/function to argue for brain ab/normality
SPECT	SPECT	Detection of gamma rays from radioisotopes in the bloodstream to image local cerebral metabolism in 3D	Research: Mostly clinical research Clinical: Imaging in neurodegenerative diseases (AD, PD, and others) Legal: Demonstrating brain dys/function to argue for brain ab/normality

CAT, computed axial tomography; CT, computed tomography; fMRI, functional magnetic resonance imaging; sMRI, structural MRI; DW-MRI, diffusion weighted MRI; SBM, surface-based morphometry; VBM, voxel-based morphology; VBLM, voxel-based lesion mapping; VBSLM, voxel-based symptom-lesion mapping; PET, positron emission imaging; SPECT, photon emission computed tomography; AD, Alzheimer's disease; PD, Parkinson's disease.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

into the specific methodological problems of neuroimaging that have surfaced more recently, I will take one more detour by considering an epistemological problem inherent in all (brain) research using correlational methods.

Epistemological Aspects of Correlational Methods in Brain Research

Neuroimaging, as we have seen, is by nature a correlational method. In fMRI research, the inferential logic is such that the measured changes of local blood flow correlate with changes of electric activity at the level of neuronal populations (so-called neurovascular coupling), which are thought to signify salient changes in local and/or large-scale network activity. In a typical experiment in cognitive neuroscience, these measurements are then related to either an experimental task (e.g., motor imagery) or used for observing the so-called "resting state": the intrinsic, ongoing activity of the "default mode" brain network in a state of wakeful relaxation.[21]

From a historical perspective, neuroimaging is no exception, as all research methods in the history of human neuroscience were correlational, whether in the phrenology of Franz Josef Gall, B.F. Skinner's behavioral psychology, or in Hans Berger's electroencephalography. Various large-scale research programs, most notably the European flagship Human Brain Project (HBP) and the United States Brain Research through Advancing Innovative Neurotechnologies® (BRAIN) initiative, are currently exploring new ways of analyzing and emulating human brain function at the micro-, meso-, and macroscale. One core idea in the HBP is the concept of *neuromorphic computing*, emulating human neural networks on electronic substrates with the mutually reinforcing principle of both improving computing through better understanding of neural circuitry and better understanding the brain through improved computing. The United States-based BRAIN initiative, in turn, takes a tool-driven approach toward advancing understanding of brain function. The core idea here derives from experiences from other large-scale research programs (e.g., the Human Genome Project or the Large Hadron Collider in particle physics), that inventing and applying new tools for brain research is more likely to yield novel and breakthrough results and models than is the scaling up of existing technologies.

Irrespective of these differences, all methods in these large-scale programs for measuring brain functions in humans, such as fMRI, EEG, MEG, neuropsychology, and others, are still correlational. It would be like having a large-scale program for modelling the global weather system and only being able to indirectly measure parameters such as wind speed, the density of cloud formations, precipitation, and air and water temperature.

Where is the problem with such an approach, one might rightly ask? If these indirect measures enable meteorologists to model weather forecasts reasonably well or predict and/or track major natural catastrophes such as hurricanes or flooding, would that not be sufficient for all practical purposes? Where is the ethical, legal, and social problem with a weather model purely based on correlational rather than direct measures? From an epistemic point of view, one may say that to truly understand the weather—irrespective if such a thing exists ontologically—we would need to know the true value (and nature?) of each parameter at each point in time. With adequate measurement equipment and sufficiently powerful computing resources we might, in principle, be able to obtain this knowledge. Therefore, the correlational approach reflects practical limits in data acquisition and processing.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

The question for consideration here is whether this is true of any complex system that is, in principle, measureable and modellable with current (or potential future) scientific methods? One might also ask whether the brain, too, is such an "ordinary" complex system or whether it is something different altogether.

This relates, to some degree, to the fundamental "hard problem"[22] in epistemology and the philosophy of consciousness of the inaccessibility of first-person phenomenal experience, e.g. the redness of red, (termed "qualia") by the current scientific method. If we take a person's utterances (or other behavior) as evidence of that person's qualia, we are again using a correlational approach rather than direct "measurement." Essentially, the question is whether we are content to treat the inner workings of the brain as a black box, as long as our correlational measurements allow for the useful modelling of cognition and other states (e.g., emotion, resting-state), or not. The question of whether and the degree to which we content ourselves with this question, in research, the courtroom, or in the social realm, determines, to some degree, the impact that methodological problems in neuroimaging will have in any of those domains.

Methodological problems in neuroimaging: general remarks and common issues

When considering potential sources of data variance, systematic biases, and analysis errors in neuroimaging, one can distinguish three different levels: (1) experimental design (number of subjects, sex ratio, task, e.g.), (2) data acquisition /measurement, and (3) data analysis. Although the first level, choosing the right experimental design and recruiting the right subject, poses similar challenges for all of neuroimaging research, the sources of error at the level of data acquisition and analysis can be quite different among methods.

As I have discussed, low statistical power resulting from small sample size is a pervasive problem in experiments in psychology and (cognitive) neuroscience and in clinical studies. This problem of "power failure"[23] both increases the likelihood of producing false positive results, and also reduces the likelihood that a significant result represents a true effect. Because of the organizational efforts and costs attached to neuroimaging, most imaging studies, also, only have small sample sizes and, therefore, problems in generating enough statistical power.

Another common problem, at the level of data analysis, is that for many methods, a number of different software packages for analysis are available. The following section on specific methods references some of the research that has demonstrated how using different software for analysis on the same data set produces different results, which is a problem because there is often no clear case to for using one software package over the other.

Methodological Problems Specific to the Neuroimaging Acquisition and Analysis Method

I will now discuss the specific methodological problems of different neuroimaging methods.

CAT/CT. In spite of the meteoric rise of MRI as the dominant neuroimaging method in research and clinical neurology and psychiatry, CAT/CT is still widely used, albeit more in clinical imaging than basic research. The main source of

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

diagnostic errors is the relatively low contrast between gray and white matter in CAT/CT scans and the heterogeneity of imaging protocols across clinics and research groups, which hampers cross-study comparability of results. One research method in which CAT/CT scans are often used is for voxel-based symptom lesion mapping.

Voxel-based and surface-based morphometry and voxel-based symptom lesion mapping. Voxel-based morphometry (VBM), surface-based morphometry (SBM), and VBSLM are popular imaging methods to study structure–function relationships in the healthy and injured brain; for example, language pathology following stroke.

In a VBM study, the group analyses from structural T1 MRI images of one group of subjects (e.g., healthy persons) are compared computationally with the group analyses from another group (e.g., patients with dementia). The voxelwise comparison of gray and white matter anatomy between the groups may show, to generate an ad-hoc example, whether dementia patients on average had less gray or white matter in a particular region in the brain than healthy persons. A specific problem for MRI-based research at the level of data acquisition is the different magnetic field strengths of different scanners (e.g., 1.5 or 3 Tesla). This makes data from different scanners difficult to combine and compare.

As in most imaging studies, VBM studies in the past, and still many today, have comparatively small sample sizes, although often these are larger than in fMRI studies. VBM is the one area in neuroimaging in which larger sample sizes are comparatively easy to generate by aggregating data, because the structural scans are less heterogeneous than, for example, task-based fMRI studies. VBM should, therefore, in theory at least, have less of a problem in generating enough statistical power to obtain decent effect sizes in comparative group studies. However, when looking at 41 meta-analyses of VBM studies on anatomical abnormalities in psychiatric conditions, John P.A. Ioannidis and colleagues could identify an excess of statistically significant results—a "significance bias"—inherent in VBM research.[24]

Longitudinal research in VBM is a scientifically appealing idea, for example, when tracking changes in gray and white matter anatomy across time to study the dynamics of brain atrophy in dementia. Unfortunately, obtaining repeated scans in a particular individual will likely increase the signal-to-noise ratio (SNR), for example because of differences in the MRI scanner's setup over time (e.g., through software updates or changes in hardware conditions). Some studies have shown that changes in SNR affect the VBM analyses, which may render the results from longitudinal studies unreliable.[25]

In terms of the comparability of analysis software across studies, several studies have shown that when analyzing the same set of structural MRI images for VBM with different software, the results may differ substantially.[26,27] The same problem seems to affect surface-based morphometry, in which only the cortical surface is considered for volumetric analysis.

In a typical symptom-lesion study with brain-injured patients, behavioral measures, in turn, such as the score in a neuropsychological test, are related to voxelwise patterns of lesions in the patients' structural brain scan (CAT/CT, high-resolution T1 MRI, or diffusion-weighted MRI scans) using multivariate statistics. Importantly, models of cognitive functions and pathology derived from fMRI research inform and influence models from lesion-deficit studies and vice versa.[28] Therefore, the neuroanatomical basis of each research method needs to be

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

particularly sound and accurate. Otherwise, systematic distortions of models derived from one imaging method (e.g., fMRI) might spill over to modeling based on another method (e.g., VBSLM), and vice versa. This problem is compounded further by the fact that recently methods for anatomical assignment in lesion-deficit maps have also been called into question. Using lesion-deficit mapping in DTI data of 581 stroke patients, Mah et al. showed that the multivariate pattern analysis usually applied to these lesions results in distorted and displaced lesion-deficit maps of anatomical location.[29]

Single-photon emission tomography (SPECT) and PET. Within the radiographic methods of neuroimaging discussed here, SPECT and PET are unique in that the source of radiographic emissions lies not outside of the body, but is injected into the bloodstream via so-called tracers, i.e. radionuclides in PET and radioisotopes in SPECT. Operating with radiochemical agents poses special methodological problems, from the production of the tracers to inter-individual differences in metabolization, which introduces additional sources of variance into such experiments.

Another particular problem, more at the level of analysis than at the level of data acquisition, which had been first recognized in PET research, but will also be a preoccupation in discussing fMRI, is the necessity of using spatial filters in preprocessing the images for analysis (so-called "smoothing"). Smoothing is applied to PET or fMRI images to account for inter-individual anatomical differences between subjects and to reduce signal noise. Often, however, the resolution of the smoothing filter is much poorer than the image resolution, which results in a loss of information.[30]

fMRI. As with the other imaging methods, the many parameters and choices a researcher makes in designing, conducting, and analyzing an fMRI-experiment influences the quality of the research. At the level of experimental design, fMRI studies also usually have small sample sizes. There is some evidence demonstrating that, apart from the problems of false positive results in single studies, small sample sizes can also lead to a "selective reporting bias."[31] Selective reporting bias refers to the over-reporting of the number of suprathreshold clusters ("activation" foci in common, albeit slightly misleading, neuroimaging parlance) in studies with small sample sizes and is thus related to the "significance bias" encountered earlier.

At the level of modeling the hemodynamic signal, the BOLD signal, in fMRI, recent studies have shown that different models may produce very different task-related effect sizes.[32] In analyzing the fMRI data, the researcher also faces the problem of cross-software comparability, with different software potentially obtaining different results.[33] Within each software package, many choices on which parameters to use for preprocessing the fMRI data are available. Here too, as in PET imaging, the problem of spatial filtering arises; that is, different sizes of the smoothing filter may lead to substantially different results in analysis.[34]

In further analyzing the data, researchers then also have to decide which level of statistical rigor they want to apply to the data. Here, they face the problem that there is no optimal way to control both for type I errors (false positives) and type II errors (false negatives); for example, when setting either a "liberal," uncorrected threshold (e.g., at $p < 0.001$) at the voxel or cluster level, or a "conservative" threshold; for example, corrected for the familywise error for their fMRI analyses.[35] Another unfortunate, yet pervasive, problem arises whenever researchers use the same fMRI data set for selecting part of the data and then performing further selective analyses on these data, a much maligned practice called "double dipping."[36] In a typical fMRI study, for example, data from approximately 300,000 voxels in the brain are acquired by the scanner. For this whole-brain set of voxels, however, only the effect sizes of the significant voxels (depending on the threshold, as mentioned) are reported, which tends to inflate the average effect sizes in fMRI studies.[37]

All of these problems and biases are rather pervasive and persistent. At the same time, neuroimaging as a research tool in cognitive neuroscience is becoming more popular every year. From the early days of using fMRI to study perception (audition, vision), motor, and cognitive (e.g., language) functions, the scope has now expanded toward the study of complex phenomena such as emotion, social cognition, metacognition or decisionmaking, to name a few. As the processes that are studied move farther away from basic neurophysiological functions (such as perception), doubts have emerged about the validity and meaning of the often surprisingly high correlations between BOLD responses and multidimensional psychological concepts such as personality or complex psychiatric disorders.[38]

Hitherto, for some time, this bouquet of methodological problems in fMRI was discussed with variable vigor and no particular urgency in the neuroimaging community. The precipitating event for forcing these lingering problems to the foreground was perhaps the publication of a study by Eklund and colleagues in 2016.[39] This study highlights important problems with the validity of familywise error control at the cluster level in simulated fMRI resting-state data. In essence, the researchers demonstrated how popular computational methods of controlling familywise error across different software packages for fMRI analysis are valid at the voxel level, but may produce high rates of false positive findings (up to 70%) at the cluster level. The study quickly drew (sometimes sensationalized) headlines and discussions in the popular science and technology press (and blogosphere) as well as a heated (and ongoing) debate within the neuroimaging community on the significance of these findings for the field.[40,41] What particularly raised the temperature of the debate were some strong claims in the paper by Eklund *et al.*, such as the notion that: "These results question the validity of some 40,000 fMRI studies and may have a large impact on the interpretation of neuroimaging results."[42] As it happens, once the first dust had settled, most commentators subsequently acknowledged the nuanced reality of the findings in the paper.[43,44]

As this debate is still ongoing, it might be too early to render a final judgment; however, it will be apparent in later sections how the neuroimaging community is now actively working toward mitigating the methodological problems discussed here.

Diffusion-weighted MRI and DTI. Diffusion-weighted MRI, particularly DTI has become a popular imaging method for assessing the anatomy of large-scale white matter fiber tracts and—it is claimed—the integrity of white matter microstructure in the human brain *in vivo*. One particularly ambitious research program called "connectomics," which aims to map the full extent (and variability) of fiber tract architecture in the brain in many different species and from the micro- to the macroscale, has emerged in recent years.[45,46,47]

The measurement parameters in DTI are quite different from the T1-weighted structural MRI (sMRI) images from clinical neuroimaging and research and fMRI images, and the method therefore carries its own methodological problems. Broadly, two large sets of problems may be distinguished. (1) The many options in setting measurement parameters in each scanner combined with the lack of a common protocol. This results in a high variability in data acquisition and makes cross-study comparisons and inferences very difficult in DTI. (2) Likewise, the lack of standardized preprocessing protocols for data analysis also has a negative impact on the comparability across experiments.

Within the DTI research and development community, these basic problems are well understood, and many articles have addressed the issue and have proposed standards and guidelines for improvement.[48,49,50,51] To give an in-depth account of all the specific methodological problems here is beyond the scope of this article. I therefore encourage the reader to consult the referenced literature for advanced discussions of the details. Important for this discussion here is the fact that the parameters of measurement and methods for analysis are even less standardized than in VBM or fMRI, which makes DTI a somewhat less reliable research tool for the time being.

Summary of the Methodological Problems

In this section, I have discussed how noise, variance, and error may exert a negative influence on the quality and reproducibility of neuroimaging at the level of experimental design, data acquisition, and data analysis across a variety of methods, such as VBM, fMRI, or DTI. Problems encountered have been underpowered studies resulting from small sample sizes, the "significance bias" in VBM research and fMRI, the software comparability problems, and the recently surfaced problem of software bugs as significant sources of variance and error. In a following section, I will discuss how this *methodological crisis,* that has been highlighted (and turbocharged) by the work of Eklund and colleagues (and other studies), had, all things considered, a rather positive net effect on opening up the neuroimaging community to internal and external scrutiny. Many neuroimaging researchers now critically review methods and develop protocols and guidelines for improvement. In the meantime, I will discuss the ethical and legal challenges that may result from the methodological crisis in neuroimaging.

Ethical Implications for Research and Clinical Practice

Ethical Implications for Research in Cognitive and Clinical Neuroscience

For research in cognitive and clinical neuroscience, the following ethical implications of the methodological problems in neuroimaging seem particularly important:

For cognitive neuroscience research, the emerging methodological problems do not seem to translate immediately into serious ethical problems with respect to the autonomy of the participating research subjects, for example concerning informed consent.

The methodological problems may, however, lead to (or perpetuate) flawed models of human brain function. As has been discussed, exaggerated effect sizes and false positive are major problems for neuroimaging. There is a real danger

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

that these methodological problems, especially when combined with overtly unethical practices such as "p-hacking" and "HARKing," may result in a growing pile of fundamentally faulty models of human cognition, behavior, or personality.

If this flawed research, however, is then used for modeling brain-based disorders—a popular research strategy in translational psychiatry and neurology—and these translational models, in turn, are used to justify particular medical interventions: the initial butterfly's flap of an obscure methodological problem may ultimately trigger a storm of, at best, useless, or, at worst, harmful interventions in patients.

Imagine, for example, an imaginary fMRI study in healthy subjects that investigates activity in the left prefrontal cortex in two groups of subjects—one scoring high on tests on impulsivity, the other scoring low—to predict behavior in average gains and losses in gambling. If the study is conducted in, for example, 30 subjects (15 per group) and the usual methods for analysis are applied, there is a high likelihood of obtaining a false positive result at the cluster level in a comparative group analysis. If the researchers were ignorant of this problem, they may translate their results into designing studies in psychiatric or neurological patients—for example, patients with chronic substance abuse, schizophrenic patients, or patients with Parkinson's disease—and, in a popular move, use reverse inferencing to conclude that altered prefrontal activity is the basis for increased impulsivity in any of these disorders. The next step in such an unfortunate self-perpetuating chain of translational research would be to apply a particular intervention, for example transcranial magnetic stimulation over the left prefrontal cortex, to modulate impulsivity in these patients.

In this example, one can see how the combination of several problems compound to first building a flawed model: the initial lack of a clear hypothesis, the statistical problems in analyzing the fMRI data, hypothesizing after the fact, and reverse inferencing. This flawed model is then used to justify particular interventions that, depending on the invasiveness and potential adverse effects (consider, e.g., the risks of deep brain stimulation), may harm subjects and/or patients. From a research ethics perspective, researchers should therefore discuss, whether the researcher has an obligation to avoid such pitfalls and to what degree he or she might be morally (and legally?) responsible for any adverse outcomes from such fundamentally flawed research. From the perspective of virtue ethics, researchers may think about how the current reality of scientific research disincentivizes researchers towards acting virtuously (i.e., follow best practices), given the constant pressure to churn out novel and exciting results in high-ranking journals at a high pace. On that note, it seems a bit discouraging that a journal's impact factor in cognitive neuroscience is correlated negatively with the average power of studies (including fMRI studies), suggesting that statistical rigor does not seem to be the highest priority in high-ranking journals when judging the quality of research.[52]

Perhaps it is the right time now to remember the previous discussion on how neuroscientific research on human brain function has, and still is, based on correlational methods. It may be asked whether this is just the way it has been—and always will be—and whether, apart from a faint feeling of epistemological unease, this poses a real problem for brain-based research? Perhaps readers will follow me in finding that in a scientific paradigm purely based on correlations, underpowered studies with unusually high effect sizes (such as in fMRI research) and spurious multivariate correlations are a particularly precarious problem. If we truly want to understand how the brain works—with forte and confidence—researchers must either find completely new ways to study the beast or safeguard our research system against these methodological problems through high-quality education and ethical adherence to best practices in neuroimaging research.

Ethical Implications for Clinical Practice

From medical history, there are many unfortunate examples of how faulty brain-based or psychological models were translated into more or less invasive, risky, and harmful interventions in patients. As an illustrative case in point, I will start with the tragic history of (pre)frontal lobotomy—or psychosurgery in general, for that matter—as a treatment for severe schizophrenia and other severe psychiatric disorders.[53]

Without the space to recount the history of lobotomy in full here—trusting, that the reader is familiar with the broad outlines—it seems worthwhile to look at the justification within the medical community for performing the procedure provided at the time. Crucially, it was not based on an up-to-date model of brain functions in the frontal lobe—which were, at least in rudimentary form, already available at the time—but rather predicated upon an idiosyncratic theory by the pioneering surgeon Egas Moniz.[54,55] The subsequent popularization of the procedure in the United States by Walter Freeman, ultimately resulting in up to 40,000 completed lobotomies, was then largely based on the practical assumption that the procedure worked, rather than systematic studies; that is, on eminence-based rather than evidence-based medicine.[56]

Another example of how the dominant *modus explanandum*, in this case psychogenic/psychoanalytic in nature, may lead to relevant suffering of vulnerable patients (and families) is provided by the "refrigerator mother theory" as the putative origin of childhood autism. First proposed (and later renounced) by psychiatrist Leo Kanner in the 1940s and then vigorously advocated for by his colleagues Bruno Bettelheim and others, the "theory" purported that autistic children had unusually cold mothers (and fathers) which—in the absence of a brain-based explanation for the condition—was considered to be a causal factor for the children's autism.[57,58]

These examples are intended to provide a small perspective on the potential negative impact of the prevailing modes of explanation, scientific paradigms, and dominant research tools at any given moment in history on the lives of vulnerable patients and their families. As researchers, we therefore have a special obligation to constantly evaluate, critically question, and improve the reliability of the tools and the validity of analytical methods that we are using in our respective fields. But what about the persuasive power of neuroimaging? What are the actual risks for patients today and the near future given the methodological problems delineated previously?

Thus far, the specific methodological problems of neuroimaging do not seem to immediately translate into concrete risk scenarios for the neurological and psychiatric patients of today. One reason is that methods such as VBM, fMRI, and DTI still have a very limited use in the clinic. Volumetric analyses of PET images (obtained with various tracers, depending on the suspected neurodegenerative disease) are used in nuclear medicine to differentiate typical from atypical Parkinson's syndromes or Alzheimer's dementia from frontotemporal dementia.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

fMRI is used in some centers for the presurgical assessment of tumor or epilepsy patients to identify functionally “eloquent” brain regions (e.g., for speech) and, still mostly in clinical research, for establishing basic communication in patients with disorders of consciousness (see Joseph J. Fins’ recent book *Rights Come to Mind* for an excellent discussion on the ethics of using fMRI (and DBS) in these patients). DTI is used for the presurgical imaging of white matter fiber tracts in patients with brain tumors. In all these scenarios, it would be difficult to construe a scenario in which the various methodological problems directly put patients at risk.

My main concern about the potential impact of the methodological problems on the clinical use of neuroimaging lies with potential applications in the near future. We are already witnessing fast progress in advanced machine learning methods, such as artificial neural networks for “deep learning,” classifying MRI scans, and predicting disease outcomes.[59,60] Such advanced machine learning algorithms can now detect morphological brain changes typical of Alzheimer’s dementia in MRI images, predict brain maturity in infants, or distinguish typical from atypical Parkinson’s syndromes, to provide a few examples.[61,62,63] Such diagnostic and predictive algorithms are very fast and efficient for finding patterns in large amounts of data, but are mostly blind toward the quality of the data itself (or methodological problems in neuroimaging analyses). Automating the diagnostic classification of neuroimages, perhaps by training a deep learning net with image data of uneven quality, or faultily modeled fMRI or DTI data analyses, could have a negative impact on the algorithmic classifications. If such fully automated image analysis system then supports physicians in clinical decisionmaking, it would be very difficult for the physicians to recognize erroneous classifications (or predictions) by the system. The responsible use of these advances in computer science—the ethics of advanced machine learning (what I would call “responsible algorithmics”)—is also a very important topic that I will explore elsewhere.

Following is a brief survey of the actual use of neuroimaging as evidence in the courtroom, and an evaluation as to whether the current methodological crisis has some bearing on this issue.

Legal Implications: Neuroimaging as Evidence in the Courtroom

In legal philosophy and “neurolaw”—the field of study linking neuroscience and legal theory and practice—there is an ongoing (and deeply entrenched) debate on the relevance of “new neuroscience” for the criminal justice system.[64] Specifically, legal scholars and researchers ask, whether current methods, conventional fMRI analyses or advanced methods for decoding brain activity, really affect the interpretation and application of *mens rea*—the concept of the “guilty mind”—as a reasonable test for criminal liability.[65] Before getting into the details of this debate, I will first survey the current use of neuroimaging as evidence in the courtroom.

The Actual Use of Neuroimaging as Evidence in United States Courtrooms

For assessing the putative impact of methodological problems in neuroimaging on legal proceedings, it might be of benefit to take a closer look at the extent to which evidence from neuroimaging is actually used in courtrooms. As no comparable in-depth studies are available for other countries thus far, I will confine my analysis here to the United States system.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

Recent excellent in-depth reviews by legal scholars have shown that neuroscience (including neuroimaging) evidence is increasingly used in United States courtrooms for adjudication. In the last 2 years alone, three comprehensive reviews have appeared that analyzed the use of neuroscience/neuroimaging evidence in United States courtrooms.[66,67,68]

As the summary chart in Table 2 illustrates, all neuroimaging modalities still play a role in United States courtrooms. The data in the reviews do not present a clear picture of the developments over time: which particular neuroimaging methods are used increasingly and which ones are on the wane. The main contexts in which neuroimaging evidence is submitted in trials, still mostly by the defense, is for assessing a defendant's guilt (guilt phase), for assessing a defendant's competency, and/or for influencing the degree of the judgment (in the penalty phase).

The following aspects seem to warrant particular considerations: (1) the courts are increasingly receptive toward accepting neuroscience data, including neuroimaging, as evidence in trials; (2) the main use of neuroimaging relates to establishing structural brain damage or brain dysfunction as mitigating factors when

Table 2. Summary of Three Comprehensive Reviews on the Use of Neuroimaging Evidence in United States Courtrooms

Study/Year	Search period	No. of cases screened	Cases involving neuroimaging* (*n*/%)
Denno 2015[66]	1992-2012	800	Total: 386/48% (of 800) CAT: 141/37% (of 386) SPECT: 27/7% (of 386) PET: 75/19% (of 386) MRI**: 143/37% (of 386)
Farahany 2016[67]	2005-2012	1585	Total: 219/14% (of 1585) *** CAT: 50/23% (of 219) SPECT: 11/5% (of 219) PET: 39/18% (of 219) sMRI: 53/24% (of 219) fMRI: 4/2% (of 219)
Gaudet and Marchant 2016[68]	?-2015	?	Total: 248/69% (of 361) **** CAT: 99/40% (of 248) SPECT: 42/17% (of 248) PET: 100/40% (of 248) sMRI: 96/39% (of 248) fMRI: 1/0.4% (of 248)

As only the review of Gaudet and Marchant provided detailed information on individual cases, the overlap of cases among the three reviews could not be determined.
*Following our earlier distinction of a narrow and broad view, neuroimaging includes all techniques from Table 1; for example, excluding EEG or MEG.
**sMRI and fMRI not differentiated
***The review included cases with EEG and unspecified methods, thus the added percentages are < 100%
****Study did not specify how many cases were screened for the 361 cases identified through text mining. As subjects often had multiple scans (e.g., CAT and MRI), the percentages for the different methods add up to more than 100%
SPECT, single-photon emission tomography; PET, positron emission tomography; CAT/CT, computed axial tomography/computed tomography; MRI, magnetic resonance imaging; sMRI, structural MRI; fMRI, functional MRI; EEG, electroencephalography; MEG, magnetoencephalography.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

considering a defendant's guilt and/or the degree of penalty; (3) far from using neuroimaging to absolve criminal defendants from guilt, the authors of the reviews summarized in Table 2 observe that, by and large, the courts use it in a modest, skilled and responsible way; and (4) there clearly is a disconnect between the media hyperbole on "brain-based lie detection" and the fact that using fMRI for such purposes is very rare in United States courtrooms.

The Influence of Neuroimagery on Judgement and Decisionmaking in the Courtroom

Apart from the factual weight of neuroimages as evidence, it should not be forgotten, however, that the persuasive imagery of brain images itself could also potentially affect decisionmaking and judgment in jurors and judges.

To date, unfortunately, only few studies have approached this question empirically and in detail.

Initial research on the general appeal of neuroscience found evidence that the "seductive allure"[69] of neuroscience-related information, particularly neuroimagery, may indeed influence judgment and decisionmaking.[70] These early studies, however, did not explicitly study these effects with respect to judgment in a legal context. Later research addressed this question by presenting neuroimages in simulated trials to participants who would be eligible as jurors. In such a study with 396 participants as potential jurors, those defendants for whom a claim for "not guilty by reason of insanity" (NGRI) was corroborated by neuroscience evidence, including neuroimaging, were more likely to receive a NGRI judgment by the jurors.[71] Not surprisingly, when considering cognitive biases such as the availability heuristic and confirmation bias, participants who reported to be influenced by neuroscience evidence were six times more likely to submit a NGRI judgment in the study.

Another study with 330 potential jurors showed that, when introducing a hypothetical scenario of fMRI-based lie detection as evidence supporting the defendants' guilt, participants were on average more likely to submit a guilty verdict.[72] The study came with an interesting twist, however, as the likelihood for a guilty verdict dropped to baseline level when fMRI-based lie detection was critically examined in cross-examination, supporting both the "allure hypothesis" and the participants' capacity for critical appraisal of the scientific evidence.

Other researchers maintain that, largely, neuroimages do not influence jurors more than oral or written testimonies on neuroscience evidence. In a study with 1,476 participants who would be eligible for jury duty, the researchers found no significant effect of neuroimages on the participants' judgments when compared with written or oral neuroscience-related evidence.[73]

For the time being, it seems that the "jury is still out" on whether the powerful allure of neuroimagery substantially influences the decisionmaking in jurors (and judges) and more research is needed on the topic.

The Effect of Neuroimaging-Based Decoding of Thoughts and Intentions for Legal Philosophy and the Courts

Detecting lies (or dishonesty or deception) is a form of decoding thoughts and intentions from brain activity. These techniques are often discussed in the popular

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

press and in neuroethics; however—as the reviews summarized in Table 2 show—are not used in the courtrooms much so far. There is a considerable debate in the neuroscience community on the plausibility and validity of current methods that claim to provide fMRI-based lie detection or the prediction of future behavior. Although appreciating the full ramification of the debate exceeds the scope of this discussion, the different positions are briefly highlighted here.

The basic claim of the proponents of neuroimaging-based lie detection is that lying (or dishonesty or deception) engages different brain networks than does telling the truth. Many fMRI experiments have found evidence in support of this hypothesis, specifically the notion that dishonesty requires increased effort of the "prefrontal control network" and honesty does not.[74,75,76,77]

Researchers who are skeptical of this claim have argued that the differences detected in some of these studies could reflect increased mental effort and not necessarily lying or dishonesty, the logic here being that simply deliberating over whether to lie or not could also increase, for example, prefrontal brain activity. Furthermore, the skeptics have pointed to the lack of replicability where, in some cases, researchers were even unable to replicate their own findings.[78] On more practical grounds, other researchers have argued that fMRI-based lie detection is unfeasible because simple countermeasures (such as slight movements or attending to nonsalient stimuli) can render an individual's fMRI session invalid or inconclusive.[79]

Implications of the Methodological Crisis for Neuroimaging-as-Evidence in the Courtroom

As the recent reviews summarized in Table 2 show, old "new neuroscience"—that is, structural and functional brain imaging to investigate for brain injury or brain dysfunction as a mitigating factor—is accepted, widely used, and mostly handled with care in the courtrooms. The numbers also show, however, that functional neuroimaging for the purpose of decoding thoughts/intentions or detecting dishonesty/deception (neuroimaging-based "lie detection") is rarely introduced (let alone admitted) in trials.

Given these considerations, what could be potential consequences of the current methodological crisis on neuroimaging-as-evidence in the courtrooms? The main reason some forms of neuroimaging (such as fMRI-based "lie detection") are rarely used in courts is a perceived lack of scientific validity; that is, current methods fail to meet the so-called Daubert standard of scientific admissibility.[80]

It is difficult to predict to what degree the current debate on the methodological problems in neuroimaging will spill over and be recognized in the legal community. In my opinion, however, the Daubert standard is where the current methodological crisis will likely have the most perceptible effect on the status of neuroimaging in the criminal justice system. Doubts on the methodological validity of neuroimaging research could quite likely raise the bar for meeting the Daubert standards, and may make courts more reluctant to admit neuroimaging as evidence into trials.

Given the rapidly progressing improvements in decoding brain states with advanced machine learning methods, however, it is important, to discuss proactively the potential impact of brain decoding on the legal system in the future. Sooner rather than later, decoding brain states will become so reliable and accurate

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

that the methods may routinely meet the Daubert standard. This, in turn, could have important consequences for legal practice and the use of neuroscience-as-evidence in the courtrooms. Changes in the ability to decode thoughts and intentions with neuroimaging (including EEG here)—new "new neuroscience"—may of course also have profound effects on the psychological and anthropological assumptions (*human nature*, to use a loaded term) underlying much of contemporary legal philosophy as well as societal attitudes toward human culpability and for justifying legal punishment.

Potential Implications of New "New Neuroscience" for Legal Philosophy and the Courtrooms

The legal scholar Stephen Morse and others have pointed out that, thus far, neuroimaging, as a paradigmatic example of "new neuroscience," does not significantly challenge the status of legal constructs (such as *mens rea*), statutes or standards in criminal justice.[81] I would tend to agree with this position, if, by "new neuroscience," what is meant is the present-day reality of neuroimaging-as-evidence in the courtrooms as described. This situation might (and will, in my opinion) change with the new "new neuroscience" of advanced machine learning for decoding intentions and thoughts from fMRI and/or EEG data (the *broad view* of neuroimaging).

Cognitive neuroscientist and philosopher Joshua Greene and others have pointed out that one should consider the possibility that new paradigms in brain decoding may indeed fundamentally change the models underlying legal concepts of criminal responsibility and culpability. In terms of prevailing theories of criminal justice, these new methods could have a transformative effect on the legal system: gradually moving the paradigm from a retributivist to a more consequentialist view of criminal justice. Some studies have already shown that brain-based explanations for criminal behavior diminish retributive attitudes toward legal punishment in participants adjudicating hypothetical cases.[82] It has been speculated whether the neurobiological explanation of this shift might be based on differences in empathy-related cognitive processing versus reward-driven processes.[83,84] Here, it is important to remember that retribution as a motivation for criminal punishment does not equal revenge, and that jurors and judges are able to acknowledge these nuances. If brain-based explanations could lead to a wider understanding of neurobiological factors in human behavior and, consequently, to more empathy, the courts may over time develop something like tender retributivism: punishment for the sake of punishing, but with a human touch.

Summary of Legal Implications of the Methodological Problems of Neuroimaging

To summarize, it seems that neuroscience and neuroimaging evidence in the courtrooms has had a positive net effect on the criminal justice system, mainly because it provides a more nuanced basis for considering brain-based mitigating factors, such as brain injury or dysfunction, in a trial. Because neuroscience-as-evidence has come to (or been summoned by) the criminal justice system to stay, a productive dialogue between researchers and legal professionals is important to ensure the responsible use, for example of neuroimaging, now and in the future. It will be interesting to see whether the coming age of brain decoding with advanced machine learning—new "new neuroscience"—will expedite *legal*

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

transformationism (á la Greene), or whether *legal immutability* (á la Morse) will prevail. The degree to which the criminal justice system will respond to and incorporate new brain-based models of human responsibility, culpability, free will, and agency will also produce repercussions on the social status of neuroscience in general and neuroimaging (and neuroimagery) in particular, which I shall investigate in depth on another occasion.

Addressing the Methodological Problems in Neuroimaging

For overcoming the crisis in reproducibility and replicability in neuroimaging (and other branches of science), I believe that the distinction between reproducible *methods* and *results* is very important and helpful. Replicating a result exactly does not itself seem like a viable end or a promising research strategy for the age of Big Data and complex systems research on the human brain.

Promoting transparency, accessibility, and consistency of the neuroimaging methods and results, however, seems to be a viable and realistic goal to weather the crisis and improve the scientificity of the field. From within the neuroimaging community, excellent proposals have already been put forth, and I encourage the reader to consult these manifestos.[85] Here, I will limit the discussion to three basic ideas: (1) promoting *best practices* in neuroimaging methodology, (2) sharing experimental designs and neuroimaging data in open science frameworks to enable replication, and (3) translational neuroimaging: analyzing large neuroimaging data sets with advanced machine learning algorithms.

Fostering Best Practices in Experimental Design, Data Acquisition, and Analysis

As I have discussed, many sources of variance and error affect the quality of neuroimaging research at different levels. At the level of experimental design, researchers have developed tools to use statistical power analysis in the design of the experiment; for example, for surface-based morphometry, to mitigate the effect of underpowered imaging studies.[86] Better yet, the scientific community should foster collaboration among research teams to increase the statistical power in neuroimaging experiments whenever possible. Importantly, a detailed and transparent report of data acquisition protocols, parameters, and methods used in data analysis (e.g., preprocessing and statistical modeling) and code should be considered mandatory, or at least a sign of quality, for publishing neuroimaging studies. The sorry state of methods reporting in published neuroimaging, as investigated by Carp et al. in 2012, attests to the urgent need to increase the transparency of neuroimaging methodology on a large scale.[87] To this end, preregistering studies and peer review on the intended study design and analysis parameters (before the first subject has seen a scanner from the inside) could also enhance the quality of neuroimaging research.

Open Neuroimaging: Sharing Data for Enabling Reproducible Research and Meta-analyses

Another approach toward increasing the reproducibility and, therefore, the validity of neuroimaging research is epitomized by the idea of open neuroimaging. Open science is a movement that is gathering significant momentum across many scientific disciplines. The basic premise is that the pervading intransparency of exact

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

methods (and code) in data analysis and proprietary publishing models compound the problem of irreproducibility. Building open databases of neuroimaging data and promoting maximum transparency (and comprehensiveness) in reporting parameter settings, analysis code, and other methods information in neuroimaging could alleviate some of the methodological problems.

The project *openfMRI* (https://openfmri.org/), for example, has now collected more than 60 fMRI data sets from more than 2,000 subjects, and many other online data repositories are now available.[88]

It should be briefly noted that storing a large amount of subject or patient data from neuroimaging in online databases may create new problems with respect to data security. Consider, for example, the increasing capability of computer scientists (and hackers) to de-anonymize personal data, in the case of online neuroimaging data, for example, by cross-referencing imaging data with information from other sources.[89] As most of these techniques also rely on advanced machine learning methods, a problematic "dual-use" scenario may occur in which advanced machine learning is beneficial for improving neuroimaging analysis and disease prediction, but is potentially harmful when it is used to expose confidential patient data. This important topic also exceeds the scope of the considerations here, but will be discussed in another article.

For now, I will briefly highlight how advanced machine learning methods might also be beneficial for overcoming some of the methodological problems in neuroimaging.

Translational Neuroimaging: Analyzing Large Neuroimaging Data Sets with Advanced Machine Learning

As I have discussed, there is still a gap between the wide use of neuroimaging methods in cognitive and clinical neuroscience research and the relatively sparse use of fMRI (or DTI) in the clinical routine of neurologists, neurosurgeons, or psychiatrists. Translational neuroimaging captures the idea of bridging this gap; for example, by leveraging the advanced machine learning methods for better pattern classification and predictive modeling of brain-based diseases.[90] Imagine the future of personalized stroke rehabilitation, for example. A translational approach would integrate the analysis and classification of multivariate data obtained from a stroke patient to model the most effective rehabilitation strategy for each patient individually. Ultimately, valid and meaningful neuroimaging data—from, for example, repeated cognitive task-based fMRI, sMRI and DTI—could play an important role in developing better predictive models and personalized therapies for a number of diseases in neurology and psychiatry.

Summary and Conclusions

I have described how the methodological crisis in neuroimaging and the vagaries attached to current methods of statistical inference in neuroimaging (particularly fMRI) bear important ethical and legal implications. For cognitive and clinical neuroscience research, I discussed how these methodological problems could create fundamentally flawed models of human cognition, behavior, and disease. This creates, from my perspective, special ethical duties for neuroimaging researchers: (1) to be fully aware and understand the nature and extent of these

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

current methodological limitations and problems; (2) to communicate accurately the uncertainties attached to some forms of neuroimaging methods to publishers and the general public (e.g., in the context of science outreach and communication activities); and (3) if feasible, to work actively toward improving on the current methodological problems; for example, by teaming up with software and data specialists and by pooling resources with other researchers

For current applications of standard neuroimaging methods in the clinical routine, the methodological problems from fMRI research currently do not seem to translate into immediate concerns for patient safety. The impending wave of translational and precision medicine, based on advanced machine learning for building personalized therapies, however, will surely leverage neuroimaging data on a large scale. I have discussed that this scenario may create a "dual-use" dynamic to this powerful technology as it potentially enables both the beneficial development of better diagnostics and therapies, but also carries the risk of misuse (e.g., for de-anonymization) and for the security of large amounts of personalized health data.

In the legal domain, old "new neuroscience," structural MRI, CAT, SPECT, and PET scans, are widely used in United States courtrooms; overwhelmingly for investigating brain injury or defects as mitigating factors regarding a criminal defendants' responsibility and/or guilt. I discussed how new "new neuroscience," specifically advances in decoding thoughts and intentions from neural data with powerful machine learning algorithms, may (or may not) have a transformative (perhaps even tendering) effect on the retributive aspects of judicial punishment.

Furthermore, I highlighted some of the laudable initiatives and efforts from researchers within the neuroimaging community for addressing the methodological problems in fMRI and other methods: (1) to develop and promote best practices and standards in data acquisition and analysis for maximizing transparency; (2) to share data in open neuroimaging frameworks; and (3) to harness the beneficial potential of advanced machine learning for improving neuroimaging data analysis.

Finally, I would like to advocate for making empirical study, as well as (neuro) ethical scholarship on the effects of the methodological crisis in neuroimaging and the future of advanced machine learning for decoding brain states integral to the funding of and education on neuroscience at any level.

Notes

1. Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences* 2016;113:7900–905.
2. Wager TD, Lindquist MA, Nichols TE, Kober H, Van Snellenberg JX. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage* 2009;45:S210–21.
3. Pauli R, Bowring A, Reynolds R, Chen G, Nichols TE, Maumet C. Exploring fMRI Results Space: 31 Variants of an fMRI Analysis in AFNI, FSL, and SPM. *Frontiers in Neuroinformatics* 2016;10:24.
4. Kazemi K, Noorizadeh N. Quantitative comparison of SPM, FSL, and Brainsuite for brain MR image segmentation. *Journal of Biomedical Physics and Engineering* 2014;4(1):13–26.
5. Lieberman MD, Cunningham WA. Type I and Type II error concerns in fMRI research: Re-balancing the scale. *Social Cognitive and Affective Neuroscience* 2009;4(4):324–8.
6. Ball T, Breckel TPK, Mutschler I, Aertsen A, Schulze-Bonhage A, Hennig J, et al. Variability of fMRI-response patterns at different spatial observation scales. *Human Brain Mapping* 2012;33:1155–1171.
7. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. *Nature Human Behaviour* 2017;1:21.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

8. Forstmeier W, Wagenmakers E-J, Parker TH. Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews* 2016 [Epub ahead of print].
9. Kerr NL. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 1998;2:196–217.
10. Gershman SJ, Daw ND. reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology* 2017;68:101–28.
11. Phelps EA, LeDoux JE. Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron* 2005;48:175–87.
12. EEG and FMRI Papers by the Numbers | Neuroscience | Human Brain Diversity Project. *Sapien Labs | Neuroscience | Human Brain Diversity Project* 2016; available at http://sapienlabs.co/500000-human-neuroscience-papers/ (last accessed 12 Mar 2017).
13. Morse S. *Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note*. Rochester, NY: Social Science Research Network; 2006.
14. Collaboration OS. Estimating the reproducibility of psychological science. *Science* 2015;349:aac4716.
15. Szucs D, Ioannidis JPA. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology* 2017;15:e2000797.
16. Begley CG, Ioannidis JPA. Reproducibility in science. *Circulation Research* 2015;116:116–26.
17. Stroebe W, Strack F. The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science* 2014;9:59–71.
18. Peng R. The reproducibility crisis in science: A statistical counterattack. *Significance* 2015;12:30–2.
19. Ashburner J, Friston KJ. Voxel-based morphometry—The methods. *NeuroImage* 2000;11:805–21.
20. Carp J. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage* 2012;63:289–300.
21. Raichle ME. The brain's default mode network. *Annual Review of Neuroscience* 2015;38:433–47.
22. Chalmers DJ. Facing up to the problem of consciousness. *Journal of Consciousness Studies* 1995;2:200–19.
23. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 2013;14:365–76.
24. Ioannidis JPA. Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry* 2011;68:773–80.
25. Shuter B, Yeh IB, Graham S, Au C, Wang S-C. Reproducibility of brain tissue volumes in longitudinal studies: Effects of changes in signal-to-noise ratio and scanner software. *NeuroImage* 2008;41:371–9.
26. Rajagopalan V, Pioro EP. Disparate voxel based morphometry (VBM) results between SPM and FSL softwares in ALS patients with frontotemporal dementia: Which VBM results to consider? *BMC Neurology* 2015;15:32.
27. Rajagopalan V, Yue GH, Pioro EP. Do preprocessing algorithms and statistical models influence voxel-based morphometry (VBM) results in amyotrophic lateral sclerosis patients? A systematic comparison of popular VBM analytical methods. *Journal of Magnetic Resonance Imaging* 2014; 40:662–7.
28. Kümmerer D, Hartwigsen G, Kellmeyer P, Glauche V, Mader I, Klöppel S, et al. Damage to ventral and dorsal language pathways in acute aphasia. *Brain* 2013;136:619–29.
29. Mah Y-H, Husain M, Rees G, Nachev P. Human brain lesion-deficit inference remapped. *Brain* 2014;137:2522–31.
30. Worsley KJ, Marrett S, Neelin P, Evans AC. Searching scale space for activation in PET images. *Human Brain Mapping* 1996;4:74–90.
31. David SP, Ware JJ, Chu IM, Loftus PD, Fusar-Poli P, Radua J, et al. Potential Reporting Bias in fMRI Studies of the Brain. *Plos One* 2013;8:e70104.
32. Lindquist MA, Meng Loh J, Atlas LY, Wager TD. Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage* 2009;45:S187–98.
33. See note 4, Kazemi, Noorizadeh 2014.
34. See note 6, Ball et al. 2012.
35. See note 5, Lieberman, Cunningham 2014
36. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience* 2009;12:535–40.
37. Reddan MC, Lindquist MA, Wager TD. Effect size estimation in neuroimaging. *JAMA Psychiatry* 2017;74:207–8.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press

38. See note 37, Reddan et al. 2017; Vul E, Harris C, Winkielman P, Pashler H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* 2009;4:274–90.
39. See note 1, Eklund et al. 2016.
40. Reynolds E. Bug in fMRI software calls 15 years of research into question. *WIRED UK*, July 6, 2016; available at http://www.wired.co.uk/article/fmri-bug-brain-scans-results (last accessed 10 Mar 2017).
41. Computer says: oops. *The Economist,* July 16, 2016; available at http://www.economist.com/news/science-and-technology/21702166-two-studies-one-neuroscience-and-one-palaeoclimatology-cast-doubt (last accessed 12 March 2017).
42. See note 1, Eklund et al. 2016, at 7900.
43. Pernet C, Nichols T. Has a software bug really called decades of brain imaging research into question? *The Guardian,* 2016; available at https://www.theguardian.com/science/head-quarters/2016/sep/30/has-a-software-bug-really-called-decades-of-brain-imaging-research-into-question (last accessed 1 Mar 2017).
44. Mumford JA, Pernet C, Yeo BTT, Nickerson D, Muhlert N, Stikov N, et al. Keep calm and scan on. *Organization for Human Brain Mapping (OHBM),* July 21, 2016; available at http://www.ohbmbrainmappingblog.com/blog/keep-calm-and-scan-on (last accessed 13 Mar 2017).
45. Zuo X-N, He Y, Betzel RF, Colcombe S, Sporns O, Milham MP. Human connectomics across the life span. *Trends in Cognitive Sciences* 2017;21:32–45.
46. Glasser MF, Smith SM, Marcus DS, Andersson JLR, Auerbach EJ, Behrens TEJ, et al. The human connectome project's neuroimaging approach. *Nature Neuroscience* 2016;19:1175–87.
47. Grayson DS, Bliss-Moreau E, Machado CJ, Bennett J, Shen K, Grant KA, et al. The rhesus monkey connectome predicts disrupted functional networks resulting from pharmacogenetic inactivation of the amygdala. *Neuron* 2016;91:453–66.
48. Guevara M, Román C, Houenou J, Duclap D, Poupon C, Mangin JF, et al. Reproducibility of superficial white matter tracts using diffusion-weighted imaging tractography. *NeuroImage* 2017;147:703–25.
49. O'Donnell LJ, Pasternak O. Does diffusion MRI tell us anything about the white matter? An overview of methods and pitfalls. *Schizophrenia Research* 2015;161:133–41.
50. Jones DK, Knösche TR, Turner R. White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *NeuroImage* 2013;73:239–54.
51. Jones DK, Symms MR, Cercignani M, Howard RJ. The effect of filter size on VBM analyses of DT-MRI data. *NeuroImage* 2005;26:546–54.
52. See note 15, Szucs, Ioannidis 2017.
53. Johnson J. A dark history: Memories of lobotomy in the new era of psychosurgery. *Medicine Studies* 2009;1:367–78.
54. Long T. Nov. 12, 1935: You Should (Not) Have a Lobotomy. *WIRED*, December 11, 2010; available at https://www.wired.com/2010/11/1112first-lobotomy/ (last accessed 14 Mar 2017).
55. Gross D, Schäfer G. Egas Moniz (1874–1955) and the "invention" of modern psychosurgery: A historical and ethical reanalysis under special consideration of Portuguese original sources. *Neurosurgical Focus* 2011;30:E8.
56. Kucharski A. History of frontal lobotomy in the United States, 1935-1955. *Neurosurgery* 1984;14:765–72.
57. Silberman S. *NeuroTribes: The Legacy of Autism and the Future of Neurodiversity*. New York: Penguin Publishing Group; 2015.
58. Harris JC. The origin and natural history of autism spectrum disorders. *Nature Neuroscience* 2016;19:1390–1.
59. Marblestone AH, Wayne G, Kording KP. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience* 2016;94.
60. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging* 2016;35:1153–9.
61. Smyser CD, Dosenbach NUF, Smyser TA, Snyder AZ, Rogers CE, Inder TE, et al. Prediction of brain maturity in infants using machine-learning algorithms. *NeuroImage* 2016;136:1–9.
62. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage* 2015;104:398–412.
63. Salvatore C, Cerasa A, Castiglioni I, Gallivanone F, Augimeri A, Lopez M, et al. Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy. *Journal of Neuroscience Methods* 2014;222:230–7.

64. Greene J, Cohen J. For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2004;359:1775–85.
65. Morse SJ. Inevitable mens rea. *Harvard Journal of Law & Public Policy* 2003;27:51.
66. Denno DW. The myth of the double-edged sword: An empirical study of neuroscience evidence in criminal cases. *Boston College Law Review* 2015;56:493.
67. Farahany NA. Neuroscience and behavioral genetics in US criminal law: An empirical analysis. *Journal of Law and the Biosciences* 2016;2:485–509.
68. Gaudet LM, Marchant GE. Under the radar: Neuroimaging evidence in the criminal courtroom. *Drake Law Review* 2016;64:577.
69. Weisberg DS, Keil FC, Goodstein J, Rawson E, Gray JR. The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience* 2007;20:470–7.
70. McCabe DP, Castel AD. Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition* 2008;107:343–52.
71. Gurley JR, Marcus DK. The effects of neuroimaging and brain injury on insanity defenses. *Behavioral Sciences & the Law* 2008;26:85–97.
72. See note 69, Weisberg et al. 2007.
73. Schweitzer NJ, Saks MJ, Murphy ER, Roskies AL, Sinnott-Armstrong W, Gaudet LM. Neuroimages as evidence in a mens rea defense: No impact. *Psychology, Public Policy, and Law* 2011;17:357–93.
74. Gaudet LM, Kerkmans JP, Anderson NE, Kiehl KA. Can neuroscience help predict future antisocial behavior. *Fordham Law Review* 2016;85:503.
75. Greene JD, Paxton JM. Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences of the United States of America* 2009; 106:12506–11.
76. Sip KE, Roepstorff A, McGregor W, Frith CD. Detecting deception: the scope and limits. *Trends in Cognitive Sciences* 2008;12:48–53.
77. Lee TMC, Liu H-L, Tan L-H, Chan CCH, Mahankali S, Feng C-M, et al. Lie detection by functional magnetic resonance imaging. *Human Brain Mapping* 2002;15:157–64.
78. Spence SA. Playing Devil's advocate†: The case against fMRI lie detection. *Legal and Criminological Psychology* 2008;13:11–25.
79. Ganis G, Rosenfeld JP, Meixner J, Kievit RA, Schendan HE. Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *NeuroImage* 2011;55:312–319.
80. Ryan MM. Daubert Standard. *LII / Legal Information Institute* 2009. Ithaca, NY: Cornell Law School, https://www.law.cornell.edu/wex/daubert_standard (last accessed March 10th 2017).
81. See note 13, Morse 2006; note 65, Morse 2003.
82. See note 64, Greene, Cohen 2004; Shariff AF, Greene JD, Karremans JC, Luguri JB, Clark CJ, Schooler JW, et al. Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological Science* 2014;25:1563–70.
83. Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, Frith CD. Empathic neural responses are modulated by the perceived fairness of others. *Nature* 2006;439:466–9.
84. Fehr E, Singer T. *The Neuroeconomics of Mind Reading and Empathy*. Rochester, NY: Social Science Research Network; 2005.
85. See note 44, Mumford et al. 2016; Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience* 2017;18:115–26; Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, et al. Best practices in data analysis and sharing in neuroimaging using MRI. *Nature Neuroscience* 2017;20:299–303.
86. Liem F, Mérillat S, Bezzola L, Hirsiger S, Philipp M, Madhyastha T, et al. Reliability and statistical power analysis of cortical and subcortical FreeSurfer metrics in a large sample of healthy elderly. *NeuroImage* 2015;108:95–109.
87. See note 20, Carp 2012.
88. Poldrack RA, Gorgolewski KJ. OpenfMRI: Open sharing of task fMRI data. *NeuroImage* 2017;144, Part B:259–61.
89. Al-Azizy D, Millard D, Symeonidis I, O'Hara K, Shadbolt N. A literature survey and classifications on data deanonymisation. In Lambrinoudakis C, Gabillon A, eds. *Risks and Security of Internet and Systems*. Cham, Switzerland: Springer International Publishing; 2015:36–51.
90. Woo C-W, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience* 2017;20:365–77.

https://doi.org/10.1017/S096318011700007X Published online by Cambridge University Press