Brain reading and the popular press

By Valtteri Arstila,
University of Turku

1. Introduction

Recent decades have not only seen neuroscience to emerge as one of the most heavily invested but also as one of the most popularized scientific disciplines. Indeed, we are constantly informed by the recent advancements and possible prospects of neuroscientific methods by articles on magazines and newspapers. From these we can learn that romance does not need to fade (Jayson, 2008), what is the difference between the democrat and republican brains (Gellene, 2007), and how wisdom manifests itself in the brain (Leake, 2009).

It comes as no surprise that the popular press takes liberties when it reports scientific studies to its readers. This practice is understandable because it is the main results of the studies and the new possibilities they suggest that matter for the layman, not their often complex details. (Moreover, most readers, if not reporters too, probably also lack the knowledge to properly understand them.) Yet this leads to an inevitable oversimplification of the studies. One outcome of this is that, without paying enough attention to the details, the results or methods of the study are taken to be more generalizable than they might actually be—researchers are claimed to be able to do (now or in the near future) things that the reported study does not warrant. In effect, the reported neuroscientific studies appear to pose more possibilities than they really have.

While the oversimplification of the scientific studies provides an overly optimistic view of the (near) future, it also raises unjustified worries: it suggests that the reported scientific studies provide means to violate our mental privacy. This makes the otherwise rather harmless generalization troublesome, especially combined with the interest shown to those studies by laymen who lack competence and knowledge to evaluate these claims critically. This calls for an investigation to what extent the real prospects of

© Res Cogitans 2011, no. 8, vol. 1, 4-24

neuroscience match with the prospects as they are pictured in the popular press. In this article, such a task is undertaken with the objective of providing a more systematic treatment of the issues at hand than what is usually given by popular press and incautious neuroscientists and neuroethicists.

In order to shed light on how the sovereignty of our mentality is threatened via the recent advancements in the brain imaging techniques, I will begin by briefly considering under what conditions and requirements neuroscientific studies in general could pose a kind of threat to the privacy of our mental life as the popular press pictures they can. This part is then followed by a longer discussion on five experiments, each of which has received a great deal of attention by non-neuroscientific communities. These cases are discussed at length because otherwise their significant characteristics would remain unnoticed and this would lead, again, to oversimplification. In particular, it will be shown that the chosen examples belong to four different types of mental phenomena and that separating these types is crucial for understanding why these methods of brain imaging have been so successful. Accordingly, the careful assessment of the five cases provides both a broad perspective on the current status of brain imaging studies and highlights the threat of generalizing their results.

2. When does brain imaging threaten privacy?

There appears to be two different worries a person might have as regards being a subject in a brain imaging experiment in addition to the general worry concerning the safety of an experiment. The first one is that something about the subjects that they would like to be oblivious about is revealed to them. An analog might be useful here: it is probable that a comprehensive DNA screening for genetic diseases shows that each person has an increased risk of having or getting some disease. Yet, these threats do not materialize in most cases. However, simply knowing such increased probabilities and feeling the threat can be enough for some people to restrict their life in a way that they would not do if they had not taken the test. Sometimes more information can have hindering and damaging effects for a person's quality of life. The same threat is present with the results gained by brain imaging since 6,6% of MRI

scans show some abnormalities (Illes et al., 2004). Naturally this is likely to make people worried even though these abnormalities are often harmless.

The second worry is that in a brain imaging experiment something that a subject wants to keep as a secret could be revealed to the researchers. Again an analog to a medical examination is useful: although a person might be relaxed with his habit of using hallucinogenic drugs, he might not want his habit to be generally known. Accordingly he might not want to give a blood example, as that would reveal it, should one look for it. The worry here is then the threat of losing sovereignty; being in a situation where we cannot control what information about ourselves is given to others.[i] Such a situation might occur, for example, when testing an applicant to certain jobs or for the purpose of health insurance where we are willing to enclose some but possibly not all information, especially if this information does not relate to the work or insurance policy we are applying for (Fuchs, 2006).

In order to really pose the above-described threat to the privacy of our mental life, the brain imaging techniques and the analysis methods related to them need to be good enough to reveal secrets that a subject prefers not to disclose. Hence the question that we need to keep asking when reading about new studies is to what extent things that a person wants to keep private can justifiably be concluded merely from the brain imaging data. Yet, instead of elaborating on what this means in practice, most writers simply appear to take the transparency of mind to brain imaging techniques as a face value. The consequences of such assumption are readily seen in the following three interrelated cases.

The first one is the thought that researchers can, even in principle, determine the contents of the memories and thoughts of an unwilling subject. Yet, this appears doubtful for a very simple reason: brain imaging techniques rely on the contrasts between two different conditions. These conditions can be established, for example, by asking a subject to recall some memory and not to do it. The subsequent analysis is then based on contrasting the brain imaging data obtained from both conditions. This means that if the subject is reluctant to follow the instructions and neither recalls the memory or thinks about it also when he is not supposed to think about it, then the required

contrast cannot be established. In other words, if a person does not want to disclose certain thoughts or memories and manages not to think about them, then even if we assume that such states could be in principle determined, in practice they cannot be read from the brain imaging data.

The second is the claim that while a subject's brain imaging data is analyzed, for any reason, the analysis may reveal something unintended—something that the investigation did not look for in the first place. Such a possibility requires that the strong notion of transparency, separated from a weaker notion where only the investigated phenomenon is revealed by brain imaging techniques, holds (Arstila & Scott, forthcoming). Indeed, this has been put forward for example by Margaret Eaton and Judy Illes (2007). It is reasonable to doubt such possibility however based on the two examples above. Although our DNA and our blood samples contain an enormous amount of information, the methods to uncover that information are rather specific; in order to determine whether a person has a certain inherited disease or has been taking some doping to improve his performance, doctors need to use the methods that are specific to those cases. Brain imaging data, likewise, can contain an enormous amount of information. Yet, even if we assume that the research paradigm was suitable for investigating also the unintended phenomenon (that is, the correct contrasts were established), almost without an exception the analysis only reveals the phenomenon that it is asked to reveal. After all, the data consists simply of a great amount of numbers and you have to ask "the right questions", to use the correct analysis methods, to uncover the phenomenon that is behind those numbers. Accordingly, it is doubtful that when one investigates, say, political preferences of a person, these same methods would show that the person has recent memories of being on vacation.

Finally, the third consequence is that sometimes all the mental phenomena are treated alike without any caution. To give an example, a few years ago BBC news (2005) reported that "Scientists say they have been able to monitor people's thoughts via scans of their brains." A closer reading of the article and an examination of the two scientific studies[ii] the article refers to show that unlike the title and abstract of the BBC article claimed, the

discussed studies were not about thoughts but about visual and auditory experiences. That is, the studies focused on our phenomenal states, which traditionally have been separated from the thoughts exactly for the reason that only such perceptual states exhibit phenomenology (Chalmers, 1996)![iii]

The problem here is related to the previous one. Whereas ignoring the fact that the successful methods to uncover certain phenomenon are often very specific to the nature of that method results in the claim that some mental phenomena could be revealed unintentionally, here it leads to the assumption that one specific method could be used (even intentionally) to uncover other mental phenomena too. That is, ignoring the facts leads to unjustified and overoptimistic predictions on the general applicability of a given technique, and in some case even to patently false claims. This happens especially when someone claims that a certain successful methods to investigate, say, visual states (like above) could be successfully used to uncover those mental phenomena that are likely to have different kinds of neural correlates.

In short, it is important to keep in mind the nature of mental phenomenon that one wants to investigate, and not merely assume the transparency of all mental phenomena to brain imaging techniques.[iv] Accordingly, brain imaging different types of mental phenomena is discussed separately in the next section. It should be noted that this classification is practical and mainly based on the ways in which these phenomena are investigated with brain imaging techniques. While such classification may not be ontologically correct, this method of classification appears justified in the case at hand because the issue is exactly how one brain imaging technique may be inadequate in investigating some mental phenomena.

3. The case studies

3.1 Scanning phenomenal states

In a recent and greatly publicized study, Thomas Naselaris' from Jack Gallant's research group at UC Berkeley showed his subjects a large set of natural images while simultaneously brain imaging them with a functional Magnetic Resonance Imaging (fMRI) machine (Naselaris, Prenger, Kay,

Oliver, & Gallant, 2009). A computer program then analyzed the obtained data in relation to the information it had concerning the spatial structure and semantic properties of the used stimuli. This step consisted classical machine learning, where the computer program made predictions based on statistical distributions on the data and with the emphasis on recognizing complex patterns of activation, and then chanced its parameters when the predictions do not match with the data. After the first step, the program was given brain imaging data of images that it had not used. Based on the prior learning and the brain imaging data obtained when subjects watched these images, the program then reconstructed an image that was likely to produce the brain activation pattern in question. Significantly, the accuracy of the program was very high in this task and the reconstructed image had a spatial structure and semantic properties very similar to the used stimulus. That is, although the program had not used these stimuli in its learning, it was able to reconstruct similar stimuli merely on the basis of fMRI data.

These studies on brain imaging visual phenomenology are impressive because here the method the researchers used did not depend on the predetermined correlations between stimuli and neural activation. Instead, the correlations that resulted from the machine learning were generalized so that the program was not limited to handle only a predetermined set of stimuli. Given that our visual phenomenology is rich in a sense that usually some parts of the experiences are always new, this kind of method is required for successful brain reading of our phenomenal states—not just distinguishing them without "real" understanding what it is that is distinguished. Given that the program now reconstructed the contents of visual phenomenology from the brain imaging data, not just discriminated between the predetermined alternatives, this study can be considered a realized case of brain reading visual phenomenology.

The success of the described method may give rise to concerns whether using such methodology is unethical or can be used for unethical purposes. Fortunately, to conclude that this study warrants the kind of worries that have been expressed in popular press is too hasty. To begin with, simply being able to decode subjects' ongoing visual (or other sensory) experiences is hardly

unethical—if we want to know at which image a person is looking at a time, we can simply look at the image itself. Moreover, because subjects do not really have any control over the images shown in the experiments, and thus on the visual experiences they are having, being able to determine their experiences does not tell us anything about the subjects themselves. Accordingly, it is not obvious how brain imaging the contents of sensory experiences themselves would be a significant violation of the privacy of a person. The more relevant issue is therefore whether this methodology can be used to brain read non-sensory states.

This however does not appear to be the case. To make a long story short, the method used in this study relied in part on two factors. One of them, emphasized by the researchers as vital for the reconstructing the stimuli, was the information gained from the prior brain imaging session. Thus, if a person simply refuses to provide reliable data in the first place (either by thinking something else that he says he thinks, or refuses to spend many hours in fMRI), the program cannot do machine learning from the activation patterns.

More importantly, another crucial factor was the brain imaging of the primary visual cortex, which in turn provided information used in the reconstruction of the structural properties of the stimuli. What makes this significant for the topic at hand is the fact that this cortical area is known to have a very particular, so called retinotopical, topography.[v] One could think of it as a kind of map: when you have a normal map, you can use it to get knowledge about how different streets, towns, lakes and so forth are spatially related. With the information from the primary visual cortex, the researchers can do the same with the visual field: retinotopical topography enables researchers to classify the locations of different elements (where the borders of objects are, which parts are more luminous than others) in the visual field (Wandell, Dumoulin, & Brewer, 2007). In a nutshell, the success of being able to reconstruct images that accurately reflected the structural properties of the stimulus relied on the very particular topography of certain cortical area.

While the above point does not lessen the significance of the achievement, it makes it unlikely that the method can be used in brain imaging

mental states that are non-sensory. The simple reason for this is that only some sensory areas (vision, touch, and to some extent hearing) are organized in this spatial fashion.[vi] Or more precisely, although we do not have any idea on how higher mental functions, including thinking (or what the exact nature of thoughts is in the first place), are neurally represented, nothing at the moment suggests that they have such a spatial organization. Moreover, while, say, images of faces and houses are partially processed in the different parts of the brain, it is not obvious that such holds for the thoughts about them. On the contrary, it appears that the prefrontal lobe is equally important for entertaining all thoughts (D'Esposito et al., 1995). Hence applying this method to a higher cognitive function is likely to be unsuccessful. This obviously emphasizes the above mentioned need to distinguish thoughts from phenomenal states, and in general the need to be specific on the kind of mental states one is talking about.

In sum, we can conclude that the methods relying on the processing on the primary sensory systems (visual, auditory, somatosensory) show a great amount of promise to meet the requirements of successful brain reading the contents of phenomenal states. At the same time, the ability to reconstruct subjects' ongoing visual (or other sensory) experiences is scarcely unethical. Furthermore, the applicability of these impressive methods to non-sensory states appears very limited as they are unlikely to work due to the differences in the way different mental states are neurally represented.[vii]

3.2 Scanning thoughts and decisions we make

Unlike our phenomenal states, our cognitive states do not necessarily reflect somewhat directly the presented stimuli or the environment we are in. Instead, what we think, believe, and remember may not have any resemblance on the situation we find ourselves to be. Thus, these states of ongoing contemplation—maybe something resembling inner speech although such a metaphor may not be entirely true—are more private states than phenomenal ones in a sense that they cannot be determined by looking at our surrounding. (Obviously, this does not mean that these states would not often be prompted by the situation we find ourselves in, merely that they do not need to be.)

Furthermore, they are sustained by our own mental activity, not by external stimuli.

Given this "control" over our thoughts, opinions, and decisions, these cognitive states are the kinds of things that we might not want to share with others. For example, we might not want others to know that we are lying, thinking about how we want to leave the company, or think that the food prepared for us is tasteless. Accordingly, brain imaging these states is more likely to cause a threat to privacy than brain imaging visual experiences.

One greatly discussed series of studies, by John-Dylan Haynes' group, on brain imaging these states focuses on the decisions made by subjects. In one of his experiments, subjects' task was to decide whether they will press a button with their left hand or with the right one (Soon, Brass, Heinze, & Haynes, 2008). In another, they were asked to decide whether to subtract or add two numbers together (Haynes et al., 2007). After the analysis of brain imaging data, Haynes was able to tell with higher than chance probability which one of the two given tasks subjects had decided to conduct. Interestingly, these guesses were based on information in the brain imaging data that preceded the subjects' conscious decision by 7 or 10 seconds depending on the study. Hence Haynes was able to tell the decisions that subjects were about to make well before they actually did them.

Understandably these results drew a great deal of attention and for example in the journal called Psychiatric News Haynes' studies were reported by writing: "give a neuroscientist an MRI machine and he can tell you exactly what a research subject will do seven seconds in the future" (Levin, 2010). Patently, this is a far cry of what Haynes' study showed however (even if we assume that people cannot react to situations in less than seven seconds).

To be clear, Haynes showed that it is possible to discriminate two predetermined conditions (i.e. which one of the two options subjects will choose) on the basis of the brain imaging data obtained before the subjects made their decisions. While this is a striking result, it is important to notice that what he did not do was to provide means to determine what those conditions are themselves. That is, he was not aiming at investigating what

were the predetermined alternatives what people were contemplating but (merely) how to discriminate the decision made between them.

Given that Haynes' studies did not focus on determining the content of ongoing cognitive states, but only to distinguish it from the other contents (which were also known before), his methods do not apply to situations where the alternatives that a subject is contemplating are not known; if you put a person to a scanner without knowing the decisions he or she might be thinking, you cannot "tell what a research subject will do seven seconds in the future." As a result, these studies are not in the position of violating our privacy because in normal, everyday situations we do not make decisions based only on the predetermined choices. Instead, our mind wanders and we contemplate whether there might be other alternatives too—maybe we want to multiply or divide the numbers, or maybe we do not even want to do the task at all! It should be noted that Haynes himself does not suggest otherwise.

The key issue as regards brain reading thoughts is therefore whether neuroscientists can determine the content of our thoughts only based on the brain imaging data. Just like with phenomenal states, only if this can be done, can the method handle the richness and unpredictability of our thoughts. Such possibility has also been suggested by popular press recently:

> New technology unveiled by Intel Corp. can read minds. No crystal ball or tarot cards required. This software uses brain scans to figure out what you're thinking, and in tests, it was 90% accurate. (Nelson, 2010)

This article reports a study by Tom Mitchell and his colleagues (Mitchell et al., 2008). Unlike in the case mentioned above, this time the study really involved thinking and not phenomenal states. Thinking here meant silently reciting the given word in one's mind. The given words themselves were concrete nouns, and they were classified to different semantic categories motivated by "sensory-motor features in neural representations of objects". That is, what we can do with the things the nouns refer to. In this sense, celery and corn, for example, are closer to each other (both being things we can eat) than either of them are to an airplane. A computer program was given 25 such categories, and it then used data from a trillion-word text corpus to classify several dozen concrete nouns according to these given categories. In the next

step, the fMRI data related to subjects' thinking these nouns were feed into the program. After the step of machine learning, the model was able to provide predictions of fMRI activation with highly significant accuracies for over 60 nouns for which fMRI data currently exists. Importantly, it can also do this with the nouns of which brain imaging data had not feed into its database. Thus one might be tempted to conclude that this method can be used to successfully brain image more complex thoughts in the near future.

However, this would be ignoring, like the newspaper article did, all the details that make the generalization of the method difficult. For example, the cited 90 % accuracy concerned the situation, where the program needed to distinguish which one of the two possible words subjects were thinking. These words were semantically far away ('airplane' versus 'celery') and hence easier to distinguish. When the words were closer to each other ('celery' versus 'corn'), the accuracy was far worse. Moreover, the study used only concrete nouns and the method used to distinguish them was based on the classification of these terms. Given that many abstract nouns, adverbs, and adjectives do not necessarily lend themselves to be classified by the means of sensory-motor features (what would be sensory-motor feature of 'righteous', 'very', or 'moderate'?), this method cannot be extended to apply to them. Together these issues highlight how limited the *current status* of the method is: it can only separate a word that a subject is thinking when it has the alternatives beforehand, and when these alternatives are concrete noun terms, preferably of very different semantic categories.

The more significant shortcoming as regards using this method to read our thoughts *in the future* is however that the brain imaging data was gathered by asking people to repeat silently a given word. Arguably this is not how we think: silently reciting one word at a time. Rather, many argue that they think via images, and those who think with words are likely to have fast and fleeting, not necessarily well-formed and serially recited words. In fact, if we thought in a way that this study suggests, the possible existence of the language of thought Fodor proposed (1975) would be much less debated about.

What is more, even granting that this method could be used to decode (not just to distinguish) all the words a subject is reciting and that there is a

language of thought, this method still cannot read our thoughts because of the poor temporal resolution of the fMRI technique on which the method is based on. The temporal resolution is in the order of 2-3 seconds, and it cannot be much improved without decreasing the spatial resolution necessary to distinguish different spatial activation patterns for different nouns. Given that people are probably able to think at least with the same speed as they can understand audio books (where 150-160 words are read aloud in a minute), this means that people can think at least five words per two seconds. At the same time, the method was based on separating brain activity of concrete nouns. Now if a person thinks, say, three different nouns during the two seconds, the resulting activation seen in an fMRI image correlates with the neural activity related to all of these three nouns. This means that the information about the activation related to a particular noun is lost and there is no information that could be used to identify what these three words were. This problem is emphasized by the fact that neural activation in general does not fade immediately, neural activation concerning one word increases activation concerning closely related words (priming effect), and that the neurophysiological changes on which fMRI relies on are the largest only 3-5 seconds after the onset of thinking. All of this is to say that the used method cannot provide results even closely to the rate we think. Together with the fact that the way the method pictures people to think is highly questionable, it does not appear that even this most advanced method of probing our cognitive states could be developed to the extent that it could be used to violate our mental privacy.

3.3 Scanning dynamic states

One thing that the first two types of mental phenomena share is that states belonging to either of them have some content (there is something that we experience, think, believe, and so forth). In this they differ from the third class of mental phenomena, which are more like attitudes or dispositions to act in a certain way. These mental phenomena do not have a particular content, but their presence is inferred from our responses. Examples of the mental phenomena belonging to this group are political preferences, sexual

orientation, the type of love involved in a relationship, and some personal features such as being an introvert or an extrovert. Each of them appears to be more on a background of our behavior than the phenomena belonging to the previous two classes. Moreover, the phenomena in this class have a longer presence than the fleeting mental states discussed above: we do not simply cease to be, say, democrats or extroverts, when we do not entertain the thought of being one.[viii] Finally, considered from the brain imaging perspective, the phenomena here are more limited in their manifestations than in the previous cases: while a person can entertain virtually an unlimited number of different thoughts, presumably his political preferences are limited to only a few alternatives. The importance of this latter aspect is that while brain imaging the cases belonging to this group, it is enough for researchers to distinguish different alternatives from each other—reconstructing the alternatives based on the brain imaging data is not necessary.

The study I want to focus on here is greatly discussed study on unconscious racist bias by Phelps. To tell a long story short, Phelps showed that the strength of amygdala activation to black-versus-white faces correlates with two indirect (unconscious) measures of race evaluation (but not with conscious measures) for the white subjects. For example, the more active a subject's amygdala is as a consequence of being exposed to the pictures of unfamiliar black faces, the longer it takes in average for them to attribute positive adjectives to people in these pictures.[ix]

Let us now assume that we do this test and get elevated amygdala activation for a white subject. Are we then justified, even to a moderate degree, to infer that this subject has an unconscious racist bias (although he himself does not know about it)? This is hardly the case because even if we ignore the suspiciousness related to tests on unconscious racism bias (Blanton & Jaccard, 2008), there remains a small and a larger problem to justify this conclusion.

The smaller one is that the correlation between the strength of the amygdale activation and the results in the unconscious racism bias tests was not very strong (0.576). In fact, subjects with the highest test scores did not have the highest amygdale activation, and vice versa.

The larger problem is that many different kinds of stimuli activates amygdala, and thus it is far from obvious whether the increased activation for a given person is due to his unconscious racist bias or due to something else. It could be, for example, that white faces are rather neutral for the subject, but unfamiliar black faces have some positive emotional value that is also known to induce increased amygdala activation. Or it could be that for some reason (maybe due to past experiences working in charity centers in poor areas populated by Afro-Americans), the subject feels sadness when seeing the pictures; sadness, grief and despair also increase activation in amygdala (Wang, McCarthy, Song, & LaBar, 2005). Yet again it could be that due their background, white people have simply seen less black faces (in their schools, in television, etc) and thus black faces, being not as familiar as white faces in the first place, draw more attention to them than white ones. In this case, the differentiating factor would not be any emotion attributed to the faces, nor unconscious racism, but the level of attention that also influences the level of amygdala activation.

Altogether, we have here thus four different possibilities why the amygdala activation is increased: i) unconscious bias, ii) positive emotions, iii) sadness and despair, and iv) the level of attention without any emotional valence. Given how different these pictured possibilities are, it remains unjustified to conclude anything of the subjects in these studies basing solely on the increased activation in amygdala. Again, if the brain imaging data does not entitle us to draw any conclusions of the subject, then it appears that this method, and those similar to it, does not violate our mental privacy. The ability to determine other mental traits appears equally possible, but also only "to an extremely limited extent" (Farah, Smith, Gawuga, Lindsell, & Foster, 2008).

3.4 Scanning neurophysiological brain structures

The fourth and final class of mental phenomena consists of those that can, to some extent, be detected without showing subjects any stimuli (unlike in the first and third class) and without requiring subjects to do anything (unlike in the second). Furthermore, the phenomena in this group differs from all the

above in that the investigated states here are thought to be stable, and that they can be (to a very limited extent) investigated merely on the basis of the neurophysiological features of the brain.

To give a couple of examples of the phenomena in this group, it has been shown that depression correlates with small hippocampuses and also with small amygdala. Small frontal lobe on the other hand correlates with decreased level of spontaneous behavior and abnormal sexual behavior. Obviously the referred cortical areas serve many distinct functions and hence it may be hasty to make one to one predictions from the size of some area to mental functions. Not all investigated phenomena belonging to this group focus merely on certain areas. On the contrary, it has been shown, for example, that the intelligence (profile) depends on the complex and structural patterns of grey and white matter anatomy (Haier, Jung, Yeo, Head, & Alkire, 2005). The same applies to a discovery in a recent study on Autism Spectrum Disorder (ASD) by Christine Ecker and colleagues (Ecker et al., 2010).

Possibly due to this more general, yet also more demanding, approach aimed at revealing spatially distributed patterns of grey matter that have bearing on ASD, the study of Ecker et al. achieved a strikingly good accuracy of 90 percent. This, in turn, caught the attention of popular press. Ben Hirschler (2010) from Reuters wrote, for example, that "a 15-minute brain scan could in future be used to test for autism, helping doctors diagnose the complex condition more cheaply and accurately". Adam Arnold (2010) from Sky News Online reciprocated with this, and continued writing that:

> At the moment, diagnosis can be time-consuming, expensive and delay children from receiving the right help and support. The new technique uses a form of brain scan. In trials on adults, it has already proved to be 90% accurate. The method is far quicker than conventional ways of identifying autism, and up to 20 times more cost effective.

Unfortunately in the similar fashion as in the newspaper articles discussed above, here too the authors have overlooked the details and consequently painted overoptimistic pictures of the prospects. Indeed, a quick look at the details of the study shows once again why it is unwarranted to conclude that a person has some sort of Autism Spectrum disorder based on his positive

results on the test, and subsequently, why the described method cannot be (at least at the moment) used for diagnostic purposes.

The main problem concerns the statistics: what the researchers were actually able to do was *to confirm an existing* ASD diagnosis, not *diagnosing* ASD itself, with 90 percent accuracy. That is, if a person is known to have Autism Spectrum Disorder, then this method has a nine to ten chance of detecting it. At the same time this result does not say anything about "detecting" ASD on people who do not have ASD diagnosed. Indeed, a closer reading of the paper shows that this method has only 80 percent accuracy for normal population: it misdiagnoses twenty percent of normal population as having Autism Spectrum Disorder. The reason why such a low probability in this task matters is that only one person out of one hundred actually has ASD. The end result of using this method is that only 4.5 percent of all of those who the method diagnoses as having ASD in fact has it.[x] In fact, it even classified 21 percent of the subjects having ADHD as having Autism Spectrum Disorder! While this does not lessen the value of this method as a confirmation method of the previous diagnosis, at the current state it is hardly a good diagnostic method, and any strong conclusions based on it are unwarranted.

While it is likely that the accuracy of the method will improve in the future, it is unclear that it will improve so much that the 95 percent of the diagnosed to have ASD would really have it (this accuracy is considered to be a scientific proof in the courts). The main reason to be skeptical about this is the fact that it is not obvious why ASD, which itself is a spectrum of disorders, would manifest itself in the brain in a way that could not be shared with some people without ASD. After all, as mentioned above, there are other confounding factors that also depend on the structural anatomy of grey matter in certain areas, such as intelligence profile (Andreasen et al., 1993) and bipolar personality (Lochhead & Parsey, 2004). Moreover, autism is a developmental disorder. Accordingly the way it manifests itself in the brain is subject to neural plasticity and assuming some core neural representation for it seems unjustified. Then again, accepting the most likely situation that there can be overlap in the neurophysiological structures of people with and without ASD means that these neurophysiological structures cannot

determine whether a person has ASD or not—it can only provide reasons to expect one or the other. This would mean that it is unlikely that the accuracy of the method improves significantly, and then there is no justification to conclude anything about the person brain imaged on the basis of that data. In this case, the described method cannot (justifiably) be used to violate our privacy.

4. Final considerations

The purpose of this paper has been to elaborate on the question to what extent current and foreseeable methods of brain imaging in the near future can be used to violate our privacy. Instead of merely assuming that such can happen, this question was approached through the issue of to what extent it is justified to conclude something based on the brain imaging data.

Given the differences in the mental phenomena that a person might want to keep private, and in their likely manifestations in the brain, it was necessary to separate four different types of mental phenomena. These were: 1) phenomenal states, 2) cognitive states, 3) dynamic states, and 4) phenomena related to neurophysiological brain structures. This classification was in essence practical because it was based on the methods that are likely to be successful in the investigations of the phenomena in question.

One important practical difference, for example, can be seen between the first two and last two phenomena. To successfully determine the presence of certain phenomena in the latter case, it is enough for the method to be able to distinguish a limited set of possible conditions from each other. After all, a person either has, say, an unconscious racist bias or not. Likewise, there are only a few options what his sexual orientation might be. As regards the first two types of mental phenomena the method of distinguishing different options is not enough, however, for the simple reason that the number of possible states that should be distinguished is virtually unlimited. Hence, here the criteria for successful brain reading is not the capacity to distinguish a limited set of alternatives from each other, but to be able reconstruct the content of a mental state based on the brain imagining data. As the examples illustrated, such achievements are well on their way especially with

phenomenal states, yet the used method is such that it does not generalize to probing cognitive states, which in turn emphasizes the difference between these two types of mental phenomena.

What these examples highlight, and what is emphasized throughout the paper, is that the successfulness of certain methods are due to the fact that they are so very specific to the phenomenon investigated. This means that even if so successful brain reading method could be developed that we were justified to determine the existence of a certain mental phenomenon based on it, this would only make the mental phenomenon in question transparent to brain imaging techniques, not others. This is one reason not to have the kinds of worries that sometimes are put forward; being able to probe what a person sees does not mean that we can violate their privacy as regards their thoughts and memories.

Another reason to be more relaxed about brain reading is that prospects of successful brain reading even for one type to mental phenomenon are currently poor. (Excluding the probing of phenomenal states which, in turn, is not very unethical.) That is, we are not in the position where we can say what a person thinks, remembers, or decides, or whether that person feels sorry the unfamiliar Afro-Americans or has an unconscious bias towards them. More importantly, it is not obvious that we would be in such a position in the near future either. As regards dynamic states, for example, many different phenomena cause similar activation and hence strong conclusions based merely on brain imaging data are unwarranted. The same applies to determining what we are thinking (and reminiscing) and to the mental phenomena that are caused by the neurophysiological structures of our brain. Given the practical limitations of brain imaging techniques, it is far from obvious either that these difficulties can be overcome without developing new techniques—something that takes time.

All this is not to say that some transparency would not be the case in the distant future, but only that it is far from obvious that it holds for currently used techniques and especially as regards all mental phenomena. At best, what can be determined with the current methods is to confirm with high probability whether certain state of affairs holds or not when there are reasons

not related to brain imaging data to think that they could hold (like other reasons to suspect that a person has ASD). Thus, as things currently stand, the prospects of the brain imaging techniques and subsequently the worries related to them as they are pictured in the popular press are grossly exaggerated.[xi] Brain imaging techniques do not pose an imminent threat to our mental privacy.

References

Andreasen, N. C., Flaum, M., Swayze II, V., O'Leary, D. S., Alliger, R., Cohen, G., et al. (1993). Intelligence and brain structure in normal individuals. *American Journal of Psychiatry*, *150*(1), 130-134.

Arnold, A. (2010). Scan Could Diagnose Autism In 15 Minutes. Sky News Online. Retrieved from http://news.sky.com/skynews/Home/UK-News/Autism-Brain-Scan-New-Test-Could-Diagnose-Autism-In-Children-In-Just-15-Minutes/Article/201008215680740.

Arstila, V. & Scott, F. (Forthcoming). Brain reading and mental privacy. *Trames*.

BBC News. (2005). "Thoughts read" via brain scans. Retrieved from http://news.bbc.co.uk/2/hi/health/4715327.stm.

Blanton, H., & Jaccard, J. (2008). Unconscious racism: A concept in pursuit of a measure. *Annual Review of Sociology*, 34, 277-297.

Chalmers, D. (1996). *The Conscious Mind*. Oxford University Press.

D'Esposito, M., Detre, J., Alsop, D., Shin, R., Atlas, S., & Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, *378*, 279-281.

Eaton, M. L., & Illes, J. (2007). Commercializing cognitive neurotechnology—the ethical terrain. *Nature biotechnology*, 25, 393-397.

Ecker, C., Marquand, A., Mourão-Miranda, J., Johnston, P., Daly, E. M., Brammer, M. J., et al. (2010). Describing the Brain in Autism in Five Dimensions--Magnetic Resonance Imaging-Assisted Diagnosis of Autism Spectrum Disorder Using a Multiparameter Classification Approach. *Journal of Neuroscience*, 30(32), 10612-10623.

Farah, M. J., Smith, M. E., Gawuga, C., Lindsell, D., & Foster, D. (2008). Brain Imaging and Brain Privacy: A Realistic Concern? *Journal of Cognitive Neuroscience*, 119-127.

Fodor, J. (1975). *The language of thought*. Cambridge, Massachusetts: Harvard University Press.

Fuchs, T. (2006). Ethical issues in neuroscience. *Current Opinion in Psychiatry*, *19*, 600-607.

Gellene, D. (2007). Study finds left-wing brain, right-wing brain. *Los Angeles Times*. Retrieved from http://www.latimes.com/news/obituaries/la-sci-politics10sep10,1,4352129.story.

Haier, R., Jung, R., Yeo, R., Head, K., & Alkire, M. T. (2005). The neuroanatomy of general intelligence: sex matters. *NeuroImage*, 25, 320-327.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current biology : CB*, *17*(4), 323-8.

Hirschler, B. (2010). Quick brain scan could screen for autism. *Reuters*. Retrieved from http://www.reuters.com/article/idUSTRE6795I420100810.

Illes, J., Rosen, A. C., Huang, L., Goldstein, R. A., Raffin, T. A., & Swan, G. (2004). Ethical consideration of incidental findings on adult brain MRI in research. *Neurology*, *62*, 888-890.

Jayson, S. (2008). Proof's in the brain scan: Romance doesn't have to fade. *USA Today*. Retrieved from http://www.usatoday.com/news/health/2008-11-16-brain-love_N.htm.

Leake, J. (2009). Found: the brain's centre of wisdom. *The Times*. Retrieved from http://www.timesonline.co.uk/tol/news/science/article6037175.ece.

Levin, A. (2010). Opening a Window on Decision Making. *Psychiatric News*, 45(12), 14.

Lochhead, R., & Parsey, R. (2004). Regional brain gray matter volume differences in patients with bipolar disorder as assessed by optimized voxel-based morphometry. *Biological psychiatry*.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., et al. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320, 1191-1195.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. *Neuron*, *63*, 902-915.

Nelson, K. (2010). Intel debuts mind-reading brain scans in NYC. *Daily News*. Retrieved from http://www.nydailynews.com/money/2010/04/08/2010-04-08_mindreading_brain_scans_debut_in_nyc.html.

Roskies, A. (2002). Neuroethics for the new millenium. *Neuron*, *35*(1), 21-3.

Soon, C., Brass, M., Heinze, H., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543-545.

Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual Field Maps in Human Cortex. *Neuron*, *56*, 366-383.

Wang, L., McCarthy, G., Song, A. W., & LaBar, K. S. (2005). Amygdala activation to sad pictures during high-field (4 tesla) functional magnetic resonance imaging. *Emotion*, 5(1), 12-22.

---

[i] This obviously assumes that we accept that our brains define who we are to the extent that one's personality, memories, and thoughts could all be determined if one would knew all the processes in the brain. Many find such neuroessentialism intuitively plausible and it is not challenged here (Roskies, 2002).

[ii] One of them was based on a phenomenon known as binocular rivalry. In this phenomenon subjects are shown different stimuli to different eyes, and yet their perception alters between these two stimuli (they do not perceive a fused image). The stimuli used in the referred study were red and blue stripy patterns, and researchers were able to tell which one of these stimuli subjects were perceiving. The other referred study used scenes from a movie (The Good, the Bad and the Ugly), and with electrodes planted to the cortex, researchers were again able to distinguish with higher than chance accuracy which scene subjects were perceiving or hearing.

[iii] This has been under debate, however, during recent years. Nevertheless, there are two other, practical reasons to separate thoughts from phenomenal states too: I) Even if there were phenomenology related to thoughts, it would not be as specific in content as our experiences (of, say, red tomato) are. II) Presumably phenomenal states and thoughts are processed largely in the different brain areas. While the phenomenology is largely due to the processing in sensory cortex, thoughts rely more on frontal cortex.

[iv] Thus instead of generalizing directly from the success made in decoding phenomenal states to the success of decoding thoughts, the prospects of brain reading our mental life based on certain studies should be evaluated in a two-step process. In the first step, the implications of the study for (mutually) similar mental phenomena should be assessed. If those appear promising, then the second step is to assess to what extend the method could be generalized to concern other types of mental phenomena too.

[v] This means that neurons in the area laid out in a way that maps to the spatial locations in retina (the layer in the eyes where photoreceptors reside). Consequently, when an image has certain spatial structure, this structure is also reflected in how it activates photoreceptors in retina (due to optics of the eye) and this structure is later reflected in the activation pattern in the imaged brain area.

[vi] The achievements with hearing and tactile sensory system are not as impressive as with vision. Yet, temporal and spatial resolution allowing, the method might be extended to be used in these cases too. The resolution might be a problem here because other sensory systems constitute far smaller a part of cortex and thus investigating them may require techniques with better resolution than the investigation of vision does.

[vii] Spatial and temporal limitations of the brain imaging techniques may also cause additional problems. It is far more difficult to do brain reading of a phenomenon that relies on small cortical areas than of those phenomena that rely on visual cortex, which covers almost a half of the cortex.

[viii] This does not mean of course that the phenomena in the third group are necessarily stable and could not change. In fact, this possibility distinguishes them from the phenomena comprising the fourth group: mental phenomena linked to stable neurophysiological structures.

[ix] Phelps began by replicating the earlier studies on unconscious racist attitudes with the use of Implicit Association test. Her results also confirmed the old results: for white subjects it takes longer time to attribute positive adjectives to pictures of unfamiliar black subjects compared to pictures of white subjects and familiar black subjects. Although there is a great deal of controversy whether such delay can be interpreted as an indication of unconscious racist bias towards black people, for our purposes we can simply assume that this interpretation is sound. In the next step Phelps compared the activation in amygdala with the scores on the unconscious racism test. Simply put, the results showed that the strength of amygdale activation to Black-versus-White faces was correlated with the measures of unconscious race evaluation.

[x] To see this, let us assume that we test 10.000 people. One percent of them have ASD, which with 10.000 people means 100 people. With the reported 90 % accuracy, 90 out of these 100 will have a positive test result and 10 are missed. The remaining 9.900 do not have ASD, but 20 percent of them are diagnosed to have one. Accordingly, 1.980 normal people will have a positive test result. Putting these two numbers together, we get 2.070 positive test results, yet only 90 (4.5%) of them really have ASD.

[xi] Of course this does not make the debate concerning the ethical implications of the possible uses of brain imaging techniques in future unimportant. However, given the (often very) speculative nature of such debates, one should be wary when they begin to influence on the current policies of conducting brain imaging experiments. Paying attention to the details shows, for example, that much of the ethical concerns related to brain imaging derive in fact already from psychological tests where brain imaging techniques are not used (Arstila & Scott, forthcoming).