Review

Possibilities and limits of mind-reading: A neurophilosophical perspective

Kathinka Evers [a,*], Mariano Sigman [b,c]

[a] Centre for Research Ethics and Bioethics (CRB), Uppsala University, Sweden
[b] Laboratorio de Neurociencia Integrativa (LNI), Departamento de Fisica, FCEN-UBA, Buenos Aires, Argentina
[c] Universidad Torcuato Di Tella, Alte. Juan Saenz Valiente 1010, Buenos Aires C1428BIJ, Argentina

A R T I C L E   I N F O

A B S T R A C T

Access to other minds once presupposed other individuals' expressions and narrations. Today, several methods have been developed which can measure brain states relevant for assessments of mental states without 1st person overt external behavior or speech. Functional magnetic resonance imaging and trace conditioning are used clinically to identify patterns of activity in the brain that suggest the presence of consciousness in people suffering from severe consciousness disorders and methods to communicate cerebrally with patients who are motorically unable to communicate. The techniques are also used non-clinically to access subjective awareness in adults and infants. In this article we inspect technical and theoretical limits on brain–machine interface access to other minds. We argue that these techniques hold promises of important medical breakthroughs, open up new vistas of communication, and of understanding the infant mind. Yet they also give rise to ethical concerns, notably misuse as a consequence of hypes and misinterpretations.

© 2013 Elsevier Inc. All rights reserved.

Contents

* Corresponding author. Address: Centre for Research Ethics and Bioethics (CRB), Uppsala University, Box 564, SE-75122 Uppsala, Sweden. Fax: +46 18 4716675.
    E-mail address: kathinka.evers@crb.uu.se (K. Evers).

1. Introduction

In typical conditions, a subject has direct access to a subset of her or his own thoughts and experience. Access to other minds is perforce indirect, passing via that other individual's expressions and narrations. Under some aspects, knowledge of another mind might be more accurate than introspective thought; for example, when emotional involvement such as love, fear, repression, or self-deceit make it difficult for us to face or see in ourselves what may be more apparent to others (friends, family, psychotherapists), who are less directly involved. Nevertheless, knowledge of others is necessarily indirect in a manner that self-knowledge need not be. For human beings, who dispose of symbolic and recursive languages, the spoken word is primordial in gaining accurate knowledge about other minds. Until recently, these behavioral expressions, body language, and narrations exhausted our possibilities of interpreting other minds and their mental states, but that situation is now rapidly changing.

Several methods have been developed which can measure brain states relevant for assessments of mental states. These states need not be conscious to the subject being mind-read. Electroencephalography (EEG), magnetoencephalography (MEG), deep-electrode recordings and near-infrared (NIRs) and magnetic resonance imaging (MRI) have opened a door that was previously sealed to entering the minds of others, and to communication without 1st person overt external behavior or speech. Notably, functional magnetic resonance imaging is used to identify patterns of activity in the brain that suggest the presence of consciousness in people suffering from severe consciousness disorders and methods to communicate cerebrally with patients who are motorically unable to communicate. It is also used non-clinically to access conscious or non-conscious subjective awareness, e.g. perceptions, or intentions in adults and infants.

These new techniques hold promises of important medical breakthroughs and new knowledge about consciousness but also give rise to ethical concerns, e.g. in terms of health care and privacy. In this article we inspect technical and theoretical limits on the brain–machine interface access to other persons' minds. We begin by reviewing recent assessments of consciousness in patients with serious consciousness disorders, proceed to non-clinical attempts to access mental contents, and conclude with a discussion of the possibilities, promises and perils.

2. Detecting consciousness from brain activity in behaviorally noncommunicative patients

2.1. Functional magnetic resonance imaging as a measure of consciousness and a method of communication

Functional magnetic resonance imaging (fMRI) maps the blood oxygenation level dependent (BOLD) response associated with neural activation to study normal and disordered functions. Because it is non-invasive, it can be used to study also vulnerable subjects, such as children, who could not be studied by other functional brain imaging methods that use radiation. Today, the fMRI technique is widely used in clinical practice as an assessment tool; for example, to identify disease risk, e.g. the early identification of Alzheimer's Disease (Wierenga & Bondi, 2007); augment surgical planning (Bookheimer, 2007); monitor rehabilitation outcome and assist in drug development (Paulus & Stein, 2007). fMRI measurements have contributed considerably to our knowledge about abnormalities of brain response in many neuropsychiatric conditions (Brown, 2007).

fMRI can also be valuable to differentiate objectively patterns of cerebral activity in patients suffering from disorders of consciousness (DOC) (Boly et al., 2005). In this context, fMRI functions as method for assessment of consciousness in these patients, when the technique is used to investigate whether they may have preserved mental abilities that evade detection using standard clinical methods that exclusively rely on behavioral indexes.

Persistent vegetative state (PVS) is characterized by wakefulness without awareness and has been described as "one of the least understood and most ethically troublesome in modern medicine" (Owen et al., 2006). PVS patients can move, grimace and have their eyes open, but are considered to be by definition unaware of themselves or their surroundings.fMRI-studies have shown that substantial portions of the cortex may still function in PVS patients. A patient's own name was found to elicit a stronger cerebral response than unfamiliar names, suggesting that the patient could process auditory stimuli to a semantic level (Perrin et al., 2006). This observation per se does not demonstrate conscious processing since experiments in normal subjects using masked stimuli demonstrate that a stream of perceptual, semantic and motor processes can occur without awareness (Dehaene et al., 1998).

In fact, at the group level cerebral activation is usually limited to subcortical and 'low-level' primary cortical areas and is often disconnected from fronto-parietal network whose coordinated activity appears to be necessary for conscious processing (Laureys, 2005).

Stronger conclusions were drawn in a recent study where a patient who fulfilled all the diagnostic clinical criteria for a vegetative state was allegedly found able to retain the ability to understand spoken commands and to respond to them through her brain activity, rather than through external behavior, such as speech or movement. The patient's apparent decision to cooperate by imagining particular tasks when asked to do so – playing a game of tennis, or moving through the rooms of her house starting from the front door – by the authors was said to represent "a clear act of intention, which confirmed beyond any doubt that she was consciously aware of herself and her surroundings" (Owen et al., 2006: 1402). A spectacular finding suggested in a subsequent study was that one minimally conscious state (MCS) patient who was totally unable to establish any functional communication at the bedside appeared able to establish communication with the researchers reliably responding yes or no to questions using directed imagery (think about A to say yes or think about B to say no) communication device (Monti et al., 2010).

fMRI has also been used to detect cognitive and emotional processing in patients diagnosed with coma. Coma has been defined as a state of unarousable unconsciousness due to dysfunction of the brain's ascending reticular activating system (ARAS), which is responsible for arousal and the maintenance of wakefulness. fMRI used on a patient with a prolonged comatose unresponsiveness following traumatic head injury revealed cortical responses to visual, auditory and tactile stimulation. The responses varied in correlation with the level of familiarity and relevance of the stimuli: stimuli from familiar people evoked significantly stronger amygdala activation than those from unfamiliar people, and direct addressing showed the same difference compared to neutral phrases (Eickhoff et al., 2008). Similar results were obtained in a more recent study suggesting that: "activity of the language network may serve as an indicator of high-level cognition and possibly volitional processes that cannot be discerned through conventional behavioral assessment alone" (Moreno, Schiff, Giacino, Kalmar, & Hirsch, 2010).

2.2. Trace-conditioning

Even very simple and inexpensive technology, conceivable for large-scale use across the planet, can be used to investigate the possibility of conscious processing in PVS patients. Classical conditioning is a simple form of associative learning in which contingencies are established between a behaviorally important stimulus (unconditioned stimulus, UCS) and a closely paired neutral stimulus (conditioned stimulus, CS). In the trace conditioning of the eyeblink response, the CS is a tone that is presented several hundred milliseconds before the UCS, which is an air puff to the cornea. The temporal demand imposed by the silent trace interval between both stimuli has been shown to engage a broad cerebral network, including the cerebellum, neocortex and hippocampus (Clark, Manns, & Squire, 2001), and to require awareness of the contingencies between stimuli (Clark & Squire, 1998). Instead, when the UCS and CS are continuous in time, referred as delay conditioning (the term is slightly confusing but this is how it is traditionally referred and we follow this convention), blink responses are elicited despite the lack of awareness of the contingency and are observed in patients with medial temporal lobe lesions (Clark & Squire, 1998; Clark & Squire, 1999). Consistent with these results, fMRI investigations of conditioning have shown overlapping patterns of activation in regions including the anterior cingulate, medial thalamus, and visual cortex during delay and trace procedures. Instead, the hippocampus, SMA, frontal operculum, middle frontal gyrus, and inferior parietal lobule are activated during trace, but not during delay, conditioning (Knight, Cheng, Smith, Stein, & Helmstetter, 2003).

Consequently, delay conditioning has been considered a hallmark of non-declarative learning (Lavond et al., 1993) and trace learning – due to its specific dependency on conscious awareness of the contingency between stimuli (Christian & Thompson, 2003) a potential "Turing-test " of consciousness (Koch, 2004). This relies on a translational hypothesis that is discussed below in detail in its logical content. This translational hypothesis assumes that if a normal person (with no obvious pathology of consciousness) only shows trace learning when showing explicit signs of consciousness of the contingency, the same must be true for agents who cannot express consciousness explicitly. This has been used to investigate consciousness in PVS patients.

Despite being unable to report awareness explicitly, PVS patients were reported able to learn this procedure (Bekinschtein et al., 2009). Learning was specific and showed an anticipatory electromyographic response to the aversive conditioning stimulus, which was substantially stronger than to the control stimulus and was augmented as the aversive stimulus approached.

While this may be considered as a signature of consciousness, the translational argument has been debated and it is also logically possible that in these patients the same logical relation (blink only if consciousness of the contingency might not be true). The best case in favor of this lack of translation is that trace conditioning can be learnt by invertebrate sea slugs (Glanzman, 1995). In simple words, the finding that PVS patients elicit trace learning has two interpretations: (1) trace learning is an adequate test for consciousness as robustly confirmed from an investigation of healthy subjects and then, using the translation argument, PVS patients have some form of consciousness; or (2) the translation argument is wrong: this logical association may break in patients wherefore trace learning is not a good test to determine consciousness.

As we discuss below, this is a formal limitation for any signature of consciousness observed in healthy subjects and then used diagnostically in subjects who cannot indicate their own consciousness via overt behavior. In this specific case, there are several lines of evidence that suggest (probabilistically) that patients who show trace learning may have consciousness.

First, the amount of learning correlated with the degree of cortical atrophy. Secondly, and more importantly, the extent of trace learning was a better indicator of recovery than other clinical assessments (Bekinschtein et al., 2009). This is exactly

what one would like to obtain from a new measure: to show some disagreement with previous criteria (if not it is not an improvement) and perform better when there is no agreement.

Finally, there is an argument of consistency between different signatures. Assessing consciousness by trace learning ability or by fMRI measurements elicited consistent results, suggesting that individuals with DOCs might have partially preserved conscious processing, which cannot be mediated by explicit reports and is not detected by behavioral assessment (Boly et al., 2007; Owen et al., 2006). All this suggests that trace learning with high probability constitutes a good marker of consciousness in healthy subjects and also in PVS patients.

2.3. Ethical and clinical implications

2.3.1. Interpretation and diagnosis

The results of these studies can be taken to suggest that patients with DOC can retain the capacity for voluntary brain activity, follow commands, execute high-level cognitive tasks and learn procedures that are believed to require awareness of the contingency. However, it needs to be confirmed that the activated regions could not be activated without any voluntary effort of the patient with DOC, i.e. that the activation could not have been wholly automatic and unconscious (Nachev & Husain, 2007). "Inferring volition, command following, mental effort, or even awareness from brain activation patterns remains hazardous without a strict demonstration that these patterns might not be exogenously triggered in an injured brain" (Goldman, 2010: 1860).

More generally, while brain activity measures obtained can index processes such as long-range integration that are idiosyncratic of executive function (Dehaene & Changeux, 2011), they cannot say much about phenomenal consciousness or the quality and specific content of subjective experience. In fact, it may be impossible to infer the presence and content of phenomenal consciousness (Block, 1998) – the qualia, the subjective experience that is not locked to specific behaviors. Our argument is that from an ethical point of view, a patient who shows obvious signs of higher cognitive functions (which can be determined empirically as argued here) should be counted as conscious even if the patient may not have phenomenal consciousness comparable to healthy subjects. When exploring tools to assess consciousness we're not dealing with logical possibilities but with probabilities. Assigning consciousness to another agent is a probabilistic inference based on markers, which, as in any Bayesian form of inference, carry signals of different strength based on experience (priors) and the current data. As the capacity of brain–machine interface communication devices increases, this communication form should constitute a more reliable input to infer presence and should be considered equivalent to external behavioral evidence.

Often, the cues to infer the state of the mind in another agent use the principle of translation: if x denotes consciousness (through language) and each time that the agent expresses x he/she expresses y, then y constitutes a physiological signature of this state and can be considered as a measure of the state denoted by x in agents which cannot express x. An example of this principle is to use trace conditioning learning (Bekinschtein, Peeters, Shalom, & Sigman, 2011) to denote consciousness in VS patients. Normal subjects only show this form of learning (y) when expressing a verbal consciousness of the contingency (x). Hence it is assumed that the expression of trace learning (y) in patients who cannot produce deliberate speech (and hence cannot produce x) is indicative of consciousness. Note that this argument may produce obvious errors as discussed in Bekinschtein et al. (2011): moving the wings may be indicative of a state of flying in hummingbirds but this does not mean that if an animal moves it wings it necessary flies. However, this translational signature constitutes in many cases the best resources we may have to assess mental states (including consciousness) on subjects that cannot produce a direct report. Similarly, when Julian Jaynes infers the form of consciousness in the history of culture, he is using text as a fossil of thought and introspection (Diuk, Slezak, Raskovsky, Sigman, & Cecchi, 2012; Jaynes, 2000). This inference of assigning the content of thought from imperfect data (because texts have been censured, degraded, changed through time and because even the simple act of writing is a form of filtering through words and text the content of the mind) can be seen in analogy with our current effort to assign a probability that another person is conscious from noisy data (Dehaene & Changeux, 2011).

These devices mediate communication when the content of language is preserved but the motor devices or even cerebral switches to make it effective are not functional. Measures, such as the trace conditioning learning, or spontaneous activation measured in fMRI can go beyond this providing markers and signature of consciousness that do not use (even by-passing the muscle) traditional linguistic schemes. Communication accordingly plays a key role in the confirmation process.

2.3.2. Communication

Patients with DOC are unable to translate neuronal activity to external behavior, e.g. verbal communication. The studies above suggest:

(a) A method by which externally noncommunicative patients can use their residual cognitive capabilities to communicate their thoughts and feelings to those around them by modulation of their own neural activity (Owen et al., 2006).
(b) An ethical imperative to give a noncommunicative person the opportunity to reveal high-level mental activity, which in turn could indicate mental capacity (Goldman, 2010).

From an ethical point of view, if a DOC patient shows obvious signs of higher cognitive functions (something that can be checked fairly reliably), then even if that does not justify the conclusion that the patient has phenomenal consciousness comparable to healthy subjects (something that is harder to check using brain imaging), the patient should still be counted as

being conscious. Moreover, if neurotechnology has opened the door to communication with previously noncommunicative patients, to the effect that the presence of consciousness has been confirmed that was previously excluded, the care provided for these patients needs to be substantially revised.

2.3.3. Care and therapeutic interventions

The clinical implications of these studies are immense. The suggestion of awareness in a patient with DOC has radical effects on decisions concerning:

(a) The type of care that should be provided. Self-evidently, awareness in a patient necessitates entirely different services, including continuous diagnoses. Even the possibility of awareness justifies continuous diagnoses and regular checks for signs of consciousness that without obvious behavioral signs would otherwise go unnoticed. In many hospitals, it is not current practice regularly to check consciousness in patients diagnosed with coma, or persistent vegetative state, once a certain time has elapsed, but the studies described above suggest a change of practice in this regard.

(b) The benefit of therapeutic interventions. Assessment of cortical function may in the long run help to identify patients who may benefit from interventional treatment, e.g., deep-brain stimulation, DBS. DBS of the unspecific thalamocortical system through midline thalamic nuclei has been shown to have an alerting effect and increase the level of arousal defined by the ability to respond to the environment, as well as motor control in a patient who had been in a minimally conscious state for 6 years (Schiff et al., 2007). Whilst the results of this study were positive, it involved only one patient, and the authors did not consider it likely that the treatment would be useful also for patients in coma or in a persistent vegetative state. Nevertheless, if larger clinical trials confirm that DBS can be used to treat the minimally conscious state, it provides hope for a better prognosis for these patients.

(c) Decisions of euthanasia. Awareness of imagery, of learning, and of others may imply a conscious perception of pain and suffering. If this should lead to induced death is an important and difficult question that consequently arises.

(d) Given the wide public health dimensions of the problem of traumatic brain injuries, the possibilities that these studies suggest arguably present a humanitarian imperative to further investigate the state of consciousness of DOC patients (Schiff et al., 2005).

The interpretation of fMRI and the relationship between cerebral activation and consciousness have spurred a number of controversies that are to date unresolved. The mentioned studies raise important questions about the level of awareness that patients with DOC can experience and how this can be measured. They challenge the validity of behavioral indexes for discerning levels of consciousness. They challenge the care that is presently being provided for these patients. However, even if the hypotheses that DOC patients can have conscious awareness and purposeful volition are supported by the evidence presented so far and accepted as an inference to the best explanation (Owen et al., 2007), further more detailed studies must be performed in order to determine the extent of awareness in DOC patients. For this to be successful, a close collaboration between philosophy of mind, cognitive neuroscience, neurophysiology and brain imaging is essential (Sörös, 2010). It seems by all accounts clear that improved diagnostic clarity concerning consciousness disorders is called for.

3. Reading thoughts from brain activity in healthy individuals

3.1. Neurotechnological access to mental contents

Neurotechnology is also used non-clinically to enter and read the contents of the human mind via its cerebral activities. As these technical possibilities advance, the question arises: how far do they permit us to enter other minds?

There is a logical limit to this pursuit: the human mind has an essentially private sphere. An individual can in some sense and measure be known by others, but not completely. Subjectivity introduces an unknowable realm in the world of every individual, as impressions of another – and maybe of oneself – always pass through a filter of subjective interpretations. To live the experience of another is a logical impossibility. Individual experience cannot be completely shared by or conveyed to another individual. This means that individuals have a private room that cannot logically be violated. It should, however, be noted that the privacy thus entailed can still be extremely limited.

We shall here focus the discussion of how far brain reading can enable mind reading on three research areas: decoding infant minds, intentions, and vision.

3.2. Decoding the infant mind from non-invasive measures of brain activity

Every parent seeks to infer the states of an infant mind (desires, pains, thoughts, dreams) conducting informal experiments to explore these seemingly inaccessible states that cannot be expressed by formal symbolic communication.

The formal exploration of the infant mind remained for a long time exclusive to philosophical inquiry and beyond empirical and scientific determination. Today, abstract philosophical conceptions develop in synergic construction with scientific knowledge, notably through an increasing neurotechnological capacity to gain a non-invasive access to the infant brain. However, there is a crucial difference between projects of brain imaging to investigate the minds of infants versus those

of comatose patients. In both cases, brain measures can be used as a decoder to infer which aspects of the universe are observable in another's mind that cannot be expressed verbally. In both cases, observation can be directed to known markers of certain thought processes (for instance, signatures of consciousness) to investigate by translation whether such processes may be instantiated in the other's mind. The crucial difference is that, unlike comatose patients, babies' body language, gaze, gestures, are a formidable communicative niche which expresses their knowledge about the world and which can be decoded with ingenious behavioral experiments of various kinds. Still, a part of infant thought may be intangible and remains completely introspective, encapsulated in the infant mind, unexpressed by any motor effector. This observation is important when assessing the capacity of brain imaging to gain access to infant minds that cannot be inferred from pure observation of behavior.

3.2.1. Infant phoneme discrimination

Observations from brain imaging and behavior are complementary. Studies using paradigms of adaptation and behavioral responses to distracters have showed convincingly that infants can discriminate phonemes In fact phoneme discrimination extends to phonemes that are not used in their native language, an ability that regresses during the first year of life (Kuhl, Wiliams, Lacerda, Stevens, & Lindblom, 1992). This method, however, does not reveal how fast infants can detect phonetic changes, whether the brain mechanisms involved resemble those used in the adults, or the variability from trial to trial in responses which may provide an idea of the degree of fluctuations in perception of the infant mind.

The first documentation of brain responses to phoneme discrimination was performed by Ghislaine and Stanislas Dehaene. They analyzed high-density recordings of event-related potentials in 3-month-old infants listening to syllables whose first consonants differed in place of articulation (Dehaene-Lambertz & Dehaene, 1994). They presented, on each trial, a sequence of five syllables (/ba/ or /ga/). In half the trials, one syllable was repeated five times. In the other half, the syllable was repeated only four times, followed by one instance of the other syllable. Any significant difference in event-related potentials (ERPs) to repeated and deviant trials indicated that the two syllables had been discriminated. The results revealed a component localized to the temporal lobes, which indexed phonological novelty in less than 400 ms. The results also revealed that following this component there was a relatively late (700–1100 ms) frontal response to novelty, reminiscent of a similar anterior negativity, which has been observed in response to unexpected visual and auditory stimuli. This suggests that, beyond phoneme discrimination, 3-month-old infants already possess a supramodal anterior network for novelty detection that can be activated in less than one second. These results go beyond an enumeration of infants' abilities (i.e. phoneme discrimination), decomposing a complex capacity into a series of processing steps, whose duration and brain implementation can be estimated. Moreover, the dynamics and tentative localization provided by EEG is reminiscent of phoneme discrimination in adults.

Further evidence of a network language instantiated in pre-verbal infants comes from fMRI studies, which can localize brain activation close to millimeter resolution (Dehaene-Lambertz, Dehaene, & Hertz-Pannier, 2002). While this is insufficient for deciphering the neural code, it provides sufficient evidence to understand the role of different cortices in the formation of a language network. A study of the brain activity evoked by normal and reversed speech in awake and sleeping 3-month-old infants showed that the left-lateralized brain regions similar to those of adults, including the superior temporal and angular gyri, were already active in infants. Thus, precursors of adult cortical language areas are active in infants well before the onset of speech production.

A subsequent study examined the functional organization of cerebral activity in 3-month-old infants when they were listening to short-sentences in their mother tongue (Dehaene-Lambertz et al., 2006). fMRI was used to obtain the dynamics of brain activity which is typically of very poor temporal resolution (Sigman, Jobert, Lebihan, & Dehaene, 2007). The infant's network of responsive regions was then parsed into functionally distinct regions based on their speed of activation and sensitivity to sentence repetition. As in the previous study, an adult-like structure of functional MRI response delays was observed along the superior temporal regions, suggesting a hierarchical processing scheme. The fastest responses were recorded in the vicinity of Heschl's gyrus (the primary auditory cortex), whereas responses became increasingly slower toward the posterior part of the superior temporal gyrus and toward the temporal poles and inferior frontal regions (Broca's area). Activation in the Broca area increased when the sentence was repeated after a 14-s delay, suggesting the early involvement of Broca's area in verbal memory.

3.2.2. Implications

Broca's area has been shown to be involved in tasks requiring language production. In fact it constitutes one of the most famous examples of brain function localization and lesions in the Broca area systematically result in a deficit on production (but not on comprehension) of oral language. The fact that Broca's area is active in infants before the babbling stage has several consequences. First, it implies that activity in this region may drive, through interactions with the perceptual system, the learning of the complex motor sequences required for future speech production. Second, the observation of mirror neurons in the macaque ventral premotor cortex, (Kohler et al., 2002) a possible homologous to Broca's area, has raised the hypothesis of a broader function of this region in action understanding and imitation (Heiser, Iacoboni, Maeda, Marcus, & Mazziotta, 2003).

In view of these observations, activation of Broca area while 3-month infants listen to sentences suggests early common frame relating speech perception and production, possibly providing an early and covert training to the (future) speech pro-

duction apparatus. However, this extension seems impossible without a direct access to brain measures, unless Broca activity is expressed by very subtle transformation which may be a covert precursor of language.

Pre-linguistic infants show brain activity that is consistent with a very rudimentary form of language. Parents naturally help and assist their infants on tasks, which they begin to accomplish, reaching, grasping, crawling, walking, etc. Gaining awareness of aspects of the newborn and the infant mind, such as language production, numerosity, objecthood, moral and social constructs, which remain silent and opaque to the external observer may change attitudes in parent care and improve and change policies in early informal education.

3.3. Reading hidden intentions

When humans engage in goal-related activities, there is an increased activity in the prefrontal cortex. This increased level of activity could be a sign of a number of different things, such as motoric response preparation, (Haggard & Elmer, 1999) reflection on potential choices, or current intentions (Lau, Rogers, & Passingham, 2006). In a recent study, subjects were given the choice between two tasks to perform: adding or subtracting two numbers, and asked to hold onto their intention during a variable period of delay during which fMRI measurements were taken (Haynes, 2007). The activity measured in the medial and lateral regions of prefrontal cortex allowed the researchers to identify which of the two tasks the subject was intending to perform. From the choice of answers it was possible to infer which task the subject had chosen. There was no explicit instruction or behavioral response prior to the onset of the response screen, subjects responded randomly and the arrangement of numbers on the screen was also random so as to exclude that any information decoded from brain activity during the delay period would be related to covert preparations of motor responses. Several regions were found to predict intentions to perform the addition or subtraction tasks. The most accurate (71%) was the medial prefrontal cortex, which did not register intention during task execution but only during the preparation delay. Reversely, a region more superior and posterior along the medial wall registered only during the execution of the chosen task. The medial prefrontal cortex showed the same overall level of activity for the two intentions suggesting that their neural difference is not in the global levels of activity but in the detailed spatial patterns of cortical responses.

A conclusion drawn from the study was that hidden intentions and covert goals are represented by distributed patterns of activity in the prefrontal cortex, and that in humans a network of brain regions including both lateral and medial prefrontal cortex contains task-specific representations (Haynes, 2007). Hidden intentions could accordingly be detected by reading these patterns of brain activity.

A future research program is whether this binary decoding strategy can be extended to more realistic situations in which subjects opt amongst many possible, perhaps an unlimited number, of options and intentions. Moreover, the focus is entirely on conscious intentions. The next step would be to detect intentions before they reach the conscious level. The question then arises whether these increased difficulties are purely technical, or whether they reflect a fundamental problem. Is it even theoretically possible to read intentions without any limits on options, or outside the realm of conscious awareness? What would be a reason for thinking that it is, or isn't, and how could this question be decided? The difference between a merely technical and a more fundamental difficulty is complex and dynamic; and it may be subject to change over time: what is considered a fundamental problem at one time might be conceived as a technical matter at another (or vice versa) as science and technology develop (Revonsuo, 2001a, 2001b; Revonsuo, 2006).

If we had a device by which we could read the undetermined and perhaps unconscious intentions of others, we would truly be capable of advanced mind reading. Studies of visual responses are relevant to consider in that light.

3.4. Predicting visual responses to images and detecting unconscious vision

3.4.1. Predicting visual responses to images
fMRI has been used to predict responses to images, and to decipher what image of a given set a person is looking at. The first studies decoded orientation (Haynes & Rees, 2005), position (Thirion et al., 2006), and object category (Cox & Savoy, 2003) from measuring activities in the visual cortex. These first studies showed that it is possible to infer and categorize simple features in the visual scene from patterns of brain activity evoked by different kinds of images. However, they typically used quite simple stimuli and a relatively small number of images to which the subject's responses were already known. The next step was to use fMRI to predict responses to new images, where the responses were not previously established, which was done in a recent study (Kay, Naselaris, Prenger, & Gallant, 2008). In this study, a newly developed decoding method based on quantitative receptive field models that characterize the relationship between visual stimuli and fMRI activity in early visual areas succeeded in identifying with high accuracy which specific natural image an observer saw for an image chosen at random from 1000 distinct images. According to the authors, it may soon be possible to decode a person's visual experience from brain activity alone (Gallant, 2008), which "prompts the thought that it may soon be possible to decode subjective perceptual experiences such as visual imagery and dreams, an idea previously restricted to the realm of science fiction." (Editor's summary, 2008) The possibility to interpret what a person is seeing without having to select from a set of known images theoretically requires better models and measurements of brain functions than are presently available, and the question again arises whether the difficulty is technical or fundamental. One aspect of this endeavor concerns the detection of unconscious responses to images.

A new important study that was recently published made a significant step in this direction (Horikawa, Tamaki, Miyawaki, & Kamitani, 2013). Patterns of brain activity were examined during dreaming and compared to waking responses to visual stimuli. The findings suggest that the visual content of dreams is represented by the same neural substrate as observed during awake perception. A conclusion drawn by the authors is that specific visual experience during sleep is represented by brain activity patterns shared by stimulus perception, providing a means to uncover subjective contents of dreaming using objective neural measurement.

3.4.2. EEG cortically coupled computer vision for rapid image search: detecting unconscious vision

The human brain is slow compared to a digital computer in which individual transistors can switch 10^6 times faster than neurons can spike. In contrast, the human visual system is superior to computers: we are far better at recognizing objects at a glance than any computer vision system. In other words, whereas the computer functions fast, it's performance on object recognition is poor, whereas we excel at recognizing objects, say, faces, although it takes quite some time for us to do it. What if the two capacities could be combined? Can we use computers to optimize the use of available human visual processors for searching through large collections of imagery?

Optimizing image throughput is an important problem in a number of disciplines, e.g. radiology or satellite reconnaissance. In the aim of coupling human vision with computer speed, a group of researchers have developed a real-time electroencephalography (EEG)-based brain–computer interface (BCI) system for triaging imagery presented using rapid serial visual presentation (Gerson, Parra, & Sajda, 2006) According to this study, a target image in a sequence of nontarget distractor images elicits in the EEC a stereotypical spatiotemporal response, which can be detected. Because the response is stereotypical, the computer can detect a subject's recognition of a target image before the subject becomes aware of recognizing the target image. As illustration, we can imagine a BCI system operating with the aim of recognizing one specific face in a huge crowd; say, people passing through a major international airport. The subject is placed in a position to observe the faces of the people passing, and if the target face appears, the response will be read by the computer – possibly without the subject ever becoming aware of this act of recognition.

A core question prompted by each of these studies is whether the difficulties that their further developments will encounter are of a purely technical nature, or whether they reflect fundamental problems. Is it only a matter of time and technical advances before the brain can reveal the mind's innermost secrets, or are there limits in principle to that quest? The possibility of principled limitations and their distinction from technical limitations would need to be specifically motivated and spelled out. We have pointed out a logical limit with reference to Leibniz' law, but with respect to scientific theory we do not assert any specific principled limitation to mind reading. The possibility of decoding a person's subjective experience from brain activity alone theoretically requires not only better models and measurements of brain functions than are presently available, but also better models of mental contents and the many languages of thought.

4. Discussion

4.1. Stereotypical mind–brain relations

In the studies described above, the focus is on thoughts (ideas, mental contents) with clear objects, such as a specific visual image, or a mathematical task, to which related processes are *stereotypically* activated in the brain, where the term "stereotypical" refers to constancy, or perpetuation without change. The thought (e.g., the intention, or recognition) is recognizable and identifiable via its cerebral process because of this stereotypical relationship.

A thought eliciting a different cerebral process each time it occurs would presumably be illegible by neurotechnology. Not just for technical reasons but also on theoretical grounds: there could be no constant cerebral patterns to recognize it by. So, the assumption of some constancy, that we may label stereotypical, in the mind–brain relationship seems crucial to mind reading via brain measurements.

The question arises: How much of our thinking is stereotypically related to specific brain processes? All of it, or can a distinction between diversely and stereotypically related mental events be drawn? If so, what distinguishes mental contents that are thus stereotypically related from those that are not? In addition to these theoretical challenges, the practical challenge for mind reading via brain reading will be to *establish* these structures of thought/brain-processes, which requires the thought and its process to be *repeatable*.

In a strict, logical sense, each thought is unique by virtue of its distinction and can never recur. However, thoughts can be categorized as *types*, and a type of thought can have many distinct *tokens*, or occurrences. In philosophy, the type/token distinction separates the type as a concept and an abstract object from the token, which is the realization, or instantiation, or occurrence of the type. For example, the concept "cat" is a type of thought that can be instantiated, e.g., by a real cat becoming a focus of attention, or by the memory of a cat. We here use the terms types and tokens leaving open the tokens' ontological categorization as, e.g., "abstract", "real", etc. Whilst distinct occurrences, as distinct, are all essentially unique, they will also, as tokens of the same type, have some relevant similarity between them permitting their identification as "that (type of) thought". In order words, the repeatability of a thought (having "the same" thought twice) is not a question of the occurrences being logically identical or indiscernible, but *similar*. Likewise, a repeated thought eliciting "the same" cerebral process elicits not identical or indiscernible but similar processes. Presumably, both the thoughts and the cerebral

related processes must be relevantly similar i.e., similar in such a way as to allow the former to be legible via the latter. In other words: only if the cerebral processes related to the tokens of a given type of thought are relevantly similar can they theoretically permit identification of that thought.

By this argument, we arrive at the following requirement for neurotechnological mind reading:

In order for a thought to be neurotechnologically legible it must be stereotypically related to a cerebral process, of which each occurrence as such is relevantly similar to every other.

In the process of reading the contents of an individual mind via its cerebral activities, a mind–brain profile can be developed. Communication is an essential element in this process, where neurotechnology offers methods to communicate without overt external behavior. The mind might become increasingly legible in proportion to this profile becoming increasingly detailed and sophisticated. This could have important clinical use, for example, as a means to refining a method of communication for behaviorally noncommunicative patients, as described above.

The question arises, if mind–brain profiles can be developed for individuals, can they also be developed for groups? Are the neuronal structures sufficiently similar across individuals for *communities* of mind–brain profiles to be developed? If they are, how much can an individual profile help outline the profile for another individual?

The replies to these questions depend on the extent and types of neuronal variability that individuals exhibit, and on their capacity for communication. Human beings are neuronally quite dissimilar, even in the case of monozygotic twins, which raises important questions concerning how (well) we are able to communicate (Evers, 2009; Evers, Kilander, & Lindau, 2007). The uniqueness of our individual brains, (our distinction) also poses a challenge for detailed brain reading.

Interindividual brain variability is caused both by genetic variability and by idioyncrasies of epigenetic, experience dependent pruning processes. An object, say, a coffee cup will cause different patterns of activity in different brains and unless it is possible to establish an unambiguous relation between a particular cup and the corresponding activity pattern – which will not be possible in the majority of cases because of methodological limitations (one would have to explore and exclude a huge space of alternatives) – one would not be able to interpret the observed pattern. Because the global layout of networks is genetically specified and similar for brains of the same species it is possible to infer from non-invasive measurements which networks are active at a particular moment in time. If the visual areas are engaged one can assume that the subject is busy either with visual perceptual tasks or with visual imagery – and this holds for all sensory modalities. If the default network is active, it is likely that the subject is idling and concerned perhaps with introspection, if reward centers light up, one can assume that the subject either expects reward or has experienced rewarding conditions – and the same holds for aversive conditions. Often, however, one can only infer that the subject deals with emotionally loaded conditions because in most the limbic structures neurons responding in the context of aversive/rewarding conditions are intermingled and spatial resolution is too low to disentangle them.

Deciphering the semantic content of the respective processes would presuppose that one has already scanned an individual brain and established correlations between the signatures of brain activity and a particular content.

However, if with the aid of communication and neurotechnology, communities of mind–brain profiles could be developed grouping together features of relevant similarity, such as aspects of neurological disorders, this could have possible therapeutic benefits.

4.2. Conclusion: possibilities and limits

The possibilities of neurotechnological mind-reading that we have today allow access to mental states without 1st person overt external behavior or speech.

With the advancement of decoders of cerebral activity (and also of other non-cerebral markers of inner thought) it is very likely that in the near future we will see a rapid progression in the capacity to observe – without mediation of language – contents of the others' mind. As discussed above, we might be able to efficiently use a subject's cerebral cortex for rapid object recognition, even when the subject is not aware of having seen the recognized object. This may be extended as a great promise to the domain of dreams, to observe in real time the content of a visual narrative during sleep. We might be able to infer a myriad of simultaneous intentions whose deliberation process to reach explicit agency is not tangible even to the same subject. We might be able to use this technology in medical situations (most notably in VS patients) where this might be the only available tool to infer another person's will. Certainly, applications in commercial setups to control objects (games, cars, airplanes) that are currently under massive development will become more frequent and effective.

There is a logical limit to these pursuits, in that an individual cannot wholly share another's experience without merging with it. Their distinction necessarily introduces a filter, an interpretation that individuates their respective points of view. In other words, by virtue of our distinction we have a private room that cannot logically be violated.

The presence of this logical limit says nothing about the extension of our privacy, except that it isn't null. It does not exclude that our unalienable privacy may be extremely small. Moreover, it does not entail that we need have privileged access to our on experiences: the fact that there is an essential incompleteness in any other person's knowledge or experience of you does not mean that there is no, or less, incompleteness in your own self-understanding. To the contrary, we have suggested that a brain decoder may access more information about, say, the intention of a subject than that which may be simply accessed by introspection (Corallo et al., 2008, Marti et al., 2010).

The specific benefits of neurotechnological mind reading for which we argue in this article are the following:

- For a person who suffers from behavioral incapacity for communication, the prospect of neurotechnological mind reading opens up promising vistas of developing alternative methods of communication.
- The development of these techniques holds promises of important medical breakthroughs, notably improvements in the care and therapeutic interventions of patients with disorders of consciousness.
- For those – parents, pediatrics, and others – interested in understanding the infant mind, the research opens promising vistas.
- For radiology or satellite reconnaissance, notably, optimizing image throughput by coupling human vision with computer speed is a promising area of research.
- For philosophy of mind and all sciences of mind, whether they are clinically orientated or not, the research into neuro-technological mind reading is exciting and appears theoretically promising.

The development of mind reading can also be perilous, however, increasingly so if or when the techniques advance. There is, notably, a risk for misuse as a consequence of hypes, exaggerations, or misinterpretations, and a potential threat to privacy unknown in history. At present, the possibilities of neurotechnological mind reading are so rudimentary that the techniques pose threats to privacy mainly in the form of misuse, but this threat might expand and increase if the techniques are refined. In that context, the question arises: who is best placed to know what goes on in a person's mind? Who is authorized to say? Does the 1st person have privileged access, or the one to perform/interpret the cerebral measurements? Already, a person's unconscious recognition of an image can be detected. How far can that be taken? Today, at the present level of science and technology: not far. Yet in the future, if better models and measurements of brain functions and mental contents are developed, the day could come when another, with the use of neurotechnology, enters your mind further than you can yourself. Is that a threat, or a promise? How we evaluate the integrity of our mind depends in part on our trust in others and our views on society: in which society we live; and which society we want to see develop in the future.

Acknowledgments

References

Bekinschtein, T. A., Peeters, M., Shalom, D., & Sigman, M. (2011). Sea slugs, subliminal pictures, and vegetative state patients: Boundaries of consciousness in classical conditioning. *Frontiers in Psychology, 2*(337).

Bekinschtein, T. A., Shalom, D. E., Forcato, C., Herrera, M., Coleman, M. R., Manes, F. F., et al (2009). Classical conditioning in the vegetative and minimally conscious state. *Nature Reviews Neuroscience, 12*(10), 1343–1349.

Block, N. (1998). Is experiencing just representing? *Philosophy and Phenomenological Research, 58*(3), 663–670.

Boly, M., Coleman, M. R., Davis, M. H., Hampshire, A., Bor, D., Moonen, G., et al (2007). When thoughts become action: An fMRI paradigm to study volitional brain activity in noncommunicative brain-injured patients. *Neuroimage, 36*, 979–992.

Boly, M., Faymonville, M.-E., Peigneux, P., Lambermont, B., Damas, P., Del Fiore, G., et al (2005). Auditory processing in severely brain injured patients: Differences between the minimally conscious state and the persistent vegetative state. *Archives of Neurology, 61*, 233–238.

Bookheimer, S. (2007). Pre-surgical language mapping with functional magnetic resonance imaging. *Neuropsychology Review, 17*, 143–157.

Brown, G. G. (2007). Functional magnetic resonance imaging in clinical practice: Look before you leap. *Neuropsychology Review, 17*, 103–106.

Christian, K. M., & Thompson, R. F. (2003). Neural substrates of eyeblink conditioning: acquisition and retention. *Learning & memory, 10*(6), 427–455.

Clark, R. E., Manns, J. R., & Squire, L. R. (2001). Trace and delay eyeblink conditioning: Contrasting phenomena of declarative and nondeclarative memory. *Psychological Science, 12*(4), 304–308.

Clark, R. E., & Squire, L. R. (1998). Classical conditioning and brain systems: The role of awareness. *Science, 280*(5360), 77–81.

Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on Introspection Distorted Subjective Time During the Dual-Task Bottleneck. *Psychological Science, 19*(11), 1110–1117.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage, 19*, 261–270.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., et al (1998). Imaging unconscious semantic priming. *Nature, 395*(6702), 597–600.

Dehaene-Lambertz, G., & Dehaene, S. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature, 370*(6487), 292–295.

Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science, 298*(5600), 2013–2015.

Dehaene-Lambertz, G., Hertz-Pannier, L., Dubois, J., Mériaux, S., Roche, A., Sigman, M., et al (2006). Functional organization of perisylvian activation during presentation of sentences in preverbal infants. *Proceedings of the National Academy of Sciences of the United States of America, 103*(38), 14240–14245.

Diuk, C., Slezak, D. F., Raskovsky, I., Sigman, M., & Cecchi, G. A. (2012). Frontiers: A quantitative philology of introspection. *Frontiers in Integrative Neuroscience, 6*(80).

Editor's summary (2008). Reading the mind. *Nature, 452*.

Eickhoff, S. B., Dafotakis, M., Grefkes, M., Stöcker, T., Shah, N. J., Schnitzler, A., et al (2008). FMRI reveals cognitive and emotional processing in a long-term comatose patient. *Experimental Neurology, 214*, 240–246.

Evers, K. (2009). *Neuroéthique. Quand la matières'éveille*. Paris: Éditions Odile Jacob.

Evers, K., Kilander, L., & Lindau, M. (2007). Insight in frontotemporal dementia. conceptual analysis and empirical evaluation of the consensus criterion "Loss of Insight" in frontotemporal dementia. *Brain & Cognition, 63*, 13–23.

Gallant, J. L. (2008). Abstractions. *Nature, 452*, xii.

Gerson, A., Parra, L., & Sajda, P. (2006). Cortically coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14*, 174–179.

Glanzman, D. L. (1995). The cellular basis of classical conditioning in *Aplysia californica* – it's less simple than you think. *Trends Neurosci., 18*, 30–36.

Goldman, S. (2010). Can functional imaging access consciousness in noncommunicative patients? *Neurology, 75*, 1860–1861.

Haggard, P., & Elmer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research, 126*, 128–133.

Haynes, J. D. (2007). Reading hidden intentions in the human brain. *Current Biology, 17*, 323–328.

Haynes, J. D., & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Reviews Neuroscience, 8*, 686–691.

Heiser, M., Iacoboni, M., Maeda, F., Marcus, J., & Mazziotta, J. C. (2003). The essential role of Broca's area in imitation. *European Journal of Neuroscience, 17*(5), 1123–1128.

Horikawa, T., Tamaki, M., Miyawaki, Y., & Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science, 3;340*(6132), 639–642.

Jaynes, J. (2000). *The origin of consciousness in the breakdown of the bicameral mind*. Mariner Books.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature, 452*, 352–356.

Knight, D., Cheng, D. T., Smith, C. N., Stein, E. A., & Helmstetter, F. J. (2003). Neural substrates mediating human delay and trace fear conditioning. *The Journal of Neuroscience, 24*(1), 218–228.

Koch, C. (2004). *The quest for consciousness: A neurobiological approach*. Roberts and Publisher.

Kohler, E., Keysers, C., Umiltà, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297*(5582), 846–848.

Kuhl, P. K., Wiliams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606–608.

Lau, H. C., Rogers, R. D., & Passingham, R. E. (2006). On measuring the perceived onsets of spontaneous actions. *The Journal of neuroscience, 26*(27), 7265–7271.

Laureys, S. (2005). The neural correlate of (un)awareness: Lessons from the vegetative state. *Trends in Cognitive Sciences, 9*(12), 556–559.

Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition, 115*(2), 303–313.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., et al (2010). Willful modulation of brain activity in disorders of consciousness. *The New England Journal of Medicine, 362*(7), 579–589.

Moreno, D. R., Schiff, N. D., Giacino, J., Kalmar, K., & Hirsch, J. (2010). A network approach to assessing cognition in disorders of consciousness. *Neurology, 75*, 1871–1878.

Nachev, P., & Husain, M. (2007). Comment on "detecting awareness in the vegetative state". *Science, 315*, 1221.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., Jolles, D., et al (2007). Response to comments on "detecting awareness in the vegetative state". *Science, 315*, 1221.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science, 31*, 1402.

Paulus, M. P., & Stein, M. B. (2007). Role of functional magnetic resonance imaging in drug discovery. *Neuropsychology Review, 17*, 179–183.

Perrin, F., Schnakers, C., Schabus, M., Degueldre, C., Goldman, S., Brédart, S., et al (2006). Brain response to one's own name in vegetative state, minimally conscious state, and locked-in syndrome. *Archives of Neurology-Chicago, 63*, 562–569.

Revonsuo, A. (2001a). Can functional brain imaging discover consciousness in the brain? *Journal of Consciousness Studies, 8*(3), 3–23.

Revonsuo, A. (2001b). Discovering the mechanisms of consciousness. *Journal of Consciousness Studies, 8*(3), 44–50.

Revonsuo, A. (2006). *Inner presence. Consciousness as a biological phenomenon*. Cambridge, MA: MIT Press.

Schiff, N. D., Giacino, J. T., Kalmar, K., Victor, J. D., Baker, D., Gerber, M., et al (2007). Behavioral improvements with thalamic stimulation after severe traumatic brain injury. *Nature, 448*, 600–603.

Schiff, N. D., Rodriguez-Moreno, D., Kamal, A., Kim, K., Giacino, J. T., Plum, F., et al (2005). FMRI reveals large-scale network activation in minimally conscious patients. *Neurology, 64*(3), 514–523.

Sigman, M., Jobert, A., Lebihan, D., & Dehaene, S. (2007). Parsing a sequence of brain activations at psychological times using fMRI. *Neuroimage*, 655–668.

Sörös, P. (2010). The neural correlates of consciousness: Electrophysiological and neuroimaging evidence for conscious processing in vegetative state. *Clinical Neurophysiology* (published online).

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., et al (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage, 33*(4), 1104–1116.

Wierenga, C., & Bondi, M. (2007). Use of functional magnetic resonance imaging in the early identification of Alzheimer's disease. *Neuropsychology Review, 17*, 127–143.