Neuroethics (2009) 2:137–149
DOI 10.1007/s12152-009-9040-1

ORIGINAL PAPER

Brain to Computer Communication: Ethical Perspectives on Interaction Models

Guglielmo Tamburrini

Received: 4 November 2008 / Accepted: 18 February 2009 / Published online: 11 March 2009

Abstract Brain Computer Interfaces (BCIs) enable one to control peripheral ICT and robotic devices by processing brain activity on-line. The potential usefulness of BCI systems, initially demonstrated in rehabilitation medicine, is now being explored in education, entertainment, intensive workflow monitoring, security, and training. Ethical issues arising in connection with these investigations are triaged taking into account technological imminence and pervasiveness of BCI technologies. By focussing on imminent technological developments, ethical reflection is informatively grounded into realistic protocols of brain-to-computer communication. In particular, it is argued that human-machine adaptation and shared control distinctively shape autonomy and responsibility issues in current BCI interaction environments. Novel personhood issues are identified and analyzed too. These notably concern (i) the "sub-personal" use of human beings in BCI-enabled cooperative problem solving, and (ii) the pro-active protection of personal identity which BCI rehabilitation therapies may afford, in the light of so-called motor theories of thinking, for the benefit of patients affected by severe motor disabilities.

G. Tamburrini (✉)
Dipartimento di Scienze fisiche, Università di Napoli Federico II,
Complesso Universitario Monte S. Angelo, Via Cintia,
80126 Napoli, Italy
e-mail: tamburrini@na.infn.it
URL: http://people.na.infn.it/~tamburrini/index.htm

Keywords Brain-computer interfaces · BCI communication protocol · Autonomy · Responsibility · Personal identity persistence · Human-machine cooperative problem solving · Sub-personal psychology

Ethical Themes in Brain-to-Computer Communication

Brain-Computer Interfaces (BCIs from now on) enable one to control peripheral ICT and robotic devices by processing brain activity on-line. The potential usefulness of BCI-enabled brain-to-computer communication was initially demonstrated in rehabilitation medicine: severely paralysed patients, who cannot benefit from more conventional rehabilitation therapies, were able to recover some communication and motor abilities by learning to use a BCI [1, 2, 3]. BCI-actuated devices developed in research laboratories throughout the world include robotic manipulators, virtual computer keyboards, and robotic wheelchairs [4, 5, 6]. BCI research is now exploring a more comprehensive repertoire of communication and control applications for both disabled and healthy users. These applications include brain-controlled virtual simulation environments [7], computer games [8], cooperative brain-computer visual processing systems [9], and BCI-actuated robotic hands [10]—in addition to alertness detectors and neurofeedback devices that are based on BCI technologies.

Distinctive ethical issues arise in connection with both healthcare applications and other prospective uses of BCI communication technologies in education, entertainment, workplace organization, security, and training.

a. What is the binding value of informed consent and last will that locked-in patients express by means of a BCI?
b. Who is responsible for damages caused by a brain-actuated mobile robot?
c. Is human dignity jeopardized by the use of unconscious or pre-conscious brain information processing in BCI-enabled, human-machine cooperative problem solving?
d. Are worker rights threatened by the use of BCI alertness detectors in intense workflow situations?
e. Should one allow the plastic brain of young people to interact with BCI-controlled computer games?
f. Does motor and mental enhancement by brain-robot networking affect user personality and threaten personal identity persistence?

These questions illustrate a wide range of ethical issues, spanning autonomy protection and promotion, moral responsibility and liability, in addition to privacy, distributive justice, personality change and personal identity persistence. One should be careful to note, however, that only questions *a* and *b* concern *present* BCI technologies and prototype systems. Questions *c-e* arise in connection with *prospective* BCI systems that are in the purview of current BCI research objectives. And question *f* concerns a motor enhancement scenario transcending the horizon of realistic research objectives, insofar as it presupposes advances in BCI-enabled communication and control that one can hardly anticipate on the basis of state-of-art research. Here, I will selectively focus on ethical issues concerning imminent BCI developments.[1]

Triage of Ethical Themes

The triaging requirement of *technological imminence*[2] suggests the opportunity of dealing first with BCI systems for restoring communication and motor capabilities in those affected by severe motor disabilities. Ethical issues concerning these systems prominently include personal autonomy, responsibility, and liability issues—such as those exemplified by questions *a–b* above. The promotion of personal autonomy afforded by BCI systems is discussed in [11]. The focus is shifted here from the promotion to the protection of human autonomy, in the light of distinctive shared control issues arising in BCI interaction contexts. Distinctive aspects of BCI responsibility and liability issues that are addressed here arise from the fact that one is not invariably in the position to predict exactly what BCI-enabled peripheral devices will do.

The personhood issues analysed here differ from those envisaged in question *f,* insofar as the enhancement of human abilities is only marginally involved.[3] One of these issues concerns the pro-active protection of personal identity which is possibly afforded by motor rehabilitation BCI therapies: according to so-called motor theories of thinking, these therapies may contrast the decline and extinction of thinking in completely locked-in patients [3, 12]. Another personhood issue addressed here is illustrated by question *c*, concerning the use of unconscious perceptual processes in the framework of BCI-enabled brain-computer cooperative problem solving. The functional roles of human beings involved in some of these cooperative problem solving tasks are readily accounted for at the *sub-personal* level, that is, without appealing to any of their intentions, beliefs, and contents of consciousness.[4] Since human "operators" are neither required to act intentionally nor to

[1] General motives for avoiding purely speculative approaches in technoethics, that is, ethical reflection on technological scenarios that are beyond the reach of technological foresight, are discussed in [32] as far as neuroethics is concerned, and in [33] as far as nanoethics is concerned.

[2] Imminence is regarded as a chief dimension for neuro-ethical triage in [32].

[3] Accordingly, the broad philosophical context of transhumanism - in the framework of which issues of cyborg identity, rights, and responsibilities are often examined [34] - is hardly relevant here.

[4] The distinction between personal and sub-personal levels of explanation in psychology was introduced in [25]. "It is only on the personal level that explanations proceed in terms of the needs, desires, intentions and beliefs of an actor in the environment."[25], p.164. In connection with the explanation of pain states, Dennett remarks: "Since the introduction of unanalysable mental qualities leads to a premature end to explanation, we may decide that such introduction is wrong, and look for alternative modes of explanation. If we do this we must abandon the explanatory level of people and their sensations and activities and turn to the *sub-personal* level of brains and events in the nervous system." [25], p. 93, emphasis mine. For a more recent analysis of this distinction, see [26].

be aware of what is their contribution in the way of cooperative problem solving, it is appropriate to ask whether and how BCI cooperative problem solving paves the way to what I will derivatively call a sub-personal use of human beings—whereby human beings are deprived of characteristic features of human mentality, action, and personhood.

I use *social pervasiveness* as an additional triaging dimension, concentrating on the identification and analysis of ethical issues that are likely to bear on the wider classes of BCI users. The distinction between invasive and non-invasive brain signal recording methods is crucial to estimate the potential pervasiveness of BCI systems. Invasive BCI technologies involve the use of recording devices inserted inside the body—cortically implanted electrodes and electrocorticography apparatuses being notable cases in point. Non-invasive BCI systems rely on non-invasive electrophysiological techniques, such as electroencephalography (EEG), and brain imaging techniques, such as functional magnetic resonance imaging (fMRI) [13]. Most non-invasive systems, however, are currently based on the EEG, insofar as EEG performances in terms of temporal resolution and practicality of use are presently unmatched by alternative non-invasive approaches.

Invasive BCIs raise implant risk, stability, reversibility, and body integrity concerns.[5] These various issues, according to a recent assessment, prevent one from making a widespread use of rehabilitation and prosthetic devices controlled by invasive BCIs.

> [A]t present, chronic electro-based prosthetic systems are a long-term approach, with near-term applications potentially limited to only the most severely disabled patients. The transition from research to widespread use will require improving the performance-risk-cost balance by increasing overall prosthetic performance and reducing surgical risk and device cost through system integration. [14].

This view is strengthened by empirical data suggesting that subjects affected by severe motor disabilities are inclined to use non-invasive BCIs. After being informed about the advantages of invasive BCIs deriving from better signal-to-noise ratio, and the corresponding disadvantages of EEG-based BCI systems in the way of slow operation and error-prone control, 16 subjects from a group of 17 patients affected by Amyotrophic Lateral Syndrome (ALS) expressed their preference for non-invasive brain-to-computer communication [12, p.480].

In discussing the outcome of this survey, Birbaumer remarks:

> ...it is fair to conclude that non-invasive BCIs using different types of EEG signals, such as slow cortical potentials, P300 or SMR oscillations, are and will remain the method of choice for communication in paralyzed and hopefully also in completely locked-in patients with ALS and other debilitating neurological diseases. [12, p. 480]

On the basis of broad cost-benefit considerations, one is led to anticipate that healthy BCI users, if any, will be inclined to opt for non-invasive BCIs too. In the light of the triaging requirement of social pervasiveness, these various evaluations and data converge to suggest the opportunity of concentrating on the ethics of non-invasive brain-to-computer communication in general, and BCI communication based on EEG in particular.

Finally, I will appeal to *ethical novelty* as a triaging dimension, focusing on ethical issues that are more difficult to deal with on the basis of pre-existing ethical conceptual frameworks and policies. In particular, I will argue that liability issues arising in connection with the operation of brain-controlled peripheral devices are not particularly novel in this sense, whereas new conceptual and policy challenges arise in connection with BCI autonomy and personhood issues.

An Exemplary Output BCI System

By focusing on present and imminent technological developments, one can effectively ground ethical reflection into realistic BCI communication protocols. In particular, an understanding of the main functional components of non-invasive, brain-to-computer BCIs (a.k.a. non-invasive *output* BCIs) is needed to isolate

[5] Chiefly based on invasive transduction technologies are so-called *input* BCIs, which fall outside the scope of this paper. Input BCIs establish computer-to-brain communication by collecting, processing, and transmitting to the brain signals that are produced by a source external to the human body.

characteristic features of BCI-related autonomy and responsibility issues, insofar as these issues are shaped by mutual adaptation and shared control problems. The distinctive character of mutual adaptations and shared control arising in BCI-enabled Human-Computer Interaction (HCI) is aptly illustrated by reference to a specific case-study, that is, the non-invasive output BCI system developed at the IDIAP Research Institute in Switzerland [5, 6].

The IDIAP BCI takes advantage of a broad feature of neuro-cognitive processing: different mental tasks activate local cortical areas in different ways, and these differences are often reflected into characteristic features of EEG signals. Accordingly, this system attempts to identify which mental task its human user is currently executing by processing EEG signals on-line. For our present concerns, it is useful to concentrate on three main functional components of the IDIAP BCI.

A. *Signal acquisition*. Human users are directed to select a mental task to execute from some suitable triad of cognitive tasks. One of these triads includes (1) the mental rotation of a geometric figure, (2) the execution of an arithmetic operation, and (3) the imagination of some specific body movement. Task execution is voluntarily initiated and self-paced by human users, while the BCI computer monitors and pre-processes EEG traces recorded at standard scalp locations.

B. *Signal analysis*. The pre-processed EEG data are analysed with the aim of detecting patterns that are typically associated to the execution of tasks 1–3. The working of this functional component is based on a classification rule. This rule is generated by means of a machine learning technique involving a supervised learning process. The learning algorithm is provided with a training set, which is chiefly formed by pairs (x_i,y_i), where $\boldsymbol{x}_i$ is a pre-processed EEG trace and $\boldsymbol{y}_i$=1, 2, 3 is its correct classification. Drawing on training experience and background knowledge about the relationship between EEG features and mental tasks 1–3, the learning algorithm generates a procedure which, given any hitherto unseen EEG instance, assigns a probability value to each one of the possible classification hypotheses. Finally, the more probable classification hypothesis is endorsed, if a sufficiently high confidence level is reached; otherwise, the system returns an "unknown" response.

C. *Signal translation*. A table of correspondences, possibly taking into account additional features of the interaction environment, associates each genuine classification judgment to a high-level control command for some peripheral device, whose piecemeal execution is controlled by the BCI computer.[6]

The signal analysis component carries out the crucial functional mapping for BCI-enabled brain-to-computer communication by assigning a classification label to sets of EEG features. How reliable is this component in its interpretation of user intents, which manifest themselves in the rehearsal of mental tasks 1–3? This problem boils down to the problem of estimating how well will the learned classification rule perform on hitherto unseen EEG instances. One may address this reliability issue both empirically (by developing suitable testing procedures for the learned rule) and theoretically (by investigating probabilistic error bounds for the learned rule in the framework of computational learning theory). It turns out, however, that both approaches involve distinctive background assumptions about training data and target functions that are difficult to buttress when the classification of neural correlates of cognitive processing is at stake. Let's see.

Reliability of Learned Rules, Stability of Neural Signals, and the Need for Mutual Adaptation

Empirical tests carried out on the IDIAP BCI learned classification rule resulted into a percentage of errors and 'unknown' responses below 5% and 30% respectively [5, p. 248]. The significance of these tests for the purpose of providing a statistical evaluation of the reliability of the learned rule on its future uses depends on various background hypotheses,[7] which notably include the assumption that both training and test data are representative of some stochastically stable phenomenon. It turns out, however, that increased familiarity with mental tasks, in addition to pervasive contextual factors, such as mental fatigue and variable attention levels, affect the EEG correlates of cognitive processing. The stability of EEG correlates is affected by purely technical factors too, insofar as EEG cap refitting operations alter electrode

[6] See [5], for a more detailed description of these functional components.

[7] For a more general discussion of this epistemic issue in connection with the justification of inductive inference, see [35], and in connection with ethical issues concerning learning robots, see [36].

positions from one session to the next [15, p. 251]. Briefly, the significance of empirical tests that are performed to assess the reliability of learned classification rules depends on boundary conditions that are difficult to isolate and control, insofar as task execution history, current mental context, and technical set-up procedures concur to affect EEG correlates of cognitive processing.

For similar reasons, one can hardly deploy the more abstract mathematical framework of statistical learning theory [16] in order to evaluate the reliability of learned rules. Probabilistic bounds on error frequency are established in this theoretical framework under the assumption that training inputs be independently drawn from a *fixed* probability distribution. But no currently available method for setting up interaction sessions allows one to buttress this assumption, insofar as a wide variety of factors unpredictably affect the statistical distribution of relevant EEG features.

Various forms of human-computer adaptation enable one to cope with deteriorating performances of learned classification rules arising from the instability of EEG correlates of cognitive processing. To begin with, (i) off-line supervised training procedures can be iterated whenever a significant decline in machine performance is observed.[8] This machine-to-human adaptation process goes with (ii) a human-to-machine on-line adaptation process based on feedback information about the outcomes of machine classification processes. This feedback information is directly supplied by the machine; alternatively, it is indirectly inferred by human users from the observation of peripheral device behaviours. An operant conditioning process is triggered by this feedback information, which enables human users to modify EEG correlates of task execution so as to improve BCI classification performances [4]. Finally, (iii) an additional machine-to-human adaptation process was recently introduced, which is based on both positive and negative error potentials. These EEG-detectable components of Event-Related Potentials (ERPs), associated to the brain processing of errors, arise a few hundred milliseconds after presentation of perceptual stimuli which will be recognized as errors. A BCI which learns to identify error potentials can undertake a self-correcting action before the human user becomes aware of errors and, a fortiori, before she can issue error correction commands.

These interpretive, control, and mutual adaptation issues do not concern the IDIAP BCI only. EEG-based output BCI systems differ from each other in the way of communication protocols, learning rules, and the choice of EEG signal components to analyse. Their commonalities, however, usually comprise supervised learning, shared control, and mutual adaptation involving perceptual or declarative feedback. These features of BCI systems distinctively shape autonomy promotion and protection issues in BCI communication environments. Let's see.

Autonomy and Message Transmission

To begin with, let us notice that BCI brain-to-computer communication protocols require an act of *delegation*, whereby human users transfer partial control of peripheral devices to a computational system. More specifically, the human user delegates both identification of high-level action intents and control of their detailed executions to a computational system. Surrendering these aspects of action control does not necessarily result into reduced personal autonomy in view of both individualistic and relational conceptions of autonomy. According to individualistic conceptions of autonomy, acting competently from one's own desires is a crucial condition for personal autonomy, insofar as the latter requires the capability to govern oneself. And indeed, patients affected by severe motor disabilities trade-off their delegation to machines for a restored capability to act on their own desires. This form of delegation promotes autonomy according to relational conceptions of autonomy too: since the repertoire of BCI-enabled actions includes social communication, severely paralysed patients learning to use a BCI may enter a variety of interpersonal relationships, thereby securing crucial competence for autonomy.[9]

[8] The possibility of deploying on-line learning methods to deal with BCI learning problems is analyzed in [15].

[9] For an introduction to individualistic and relational conceptions of autonomy, see [37]. The promotion of autonomy of locked-in patients afforded by BCI systems is appropriately emphasized in [11], pp. 127-129. On more general grounds, however, one should carefully note, as Hansson does, that "subordination to technology will probably become an increasingly serious problem as enabling technology is developed that exhibits more and more intelligent behavior" [38, 264].

Let us now turn to consider these conditions for autonomy from an epistemic perspective. To an external observer, the operation of an EEG-based BCI provides evidence, if any, that its human user made—and is continuing to endorse—the decision to share action control with a machine. In the case of the IDIAP BCI, this evidence is strengthened by the fact that EEG signals activating peripheral devices are unlikely to be unreflectively and unintentionally produced by human users. Indeed, these signals arise in connection with mental activities which normally require sustained volitional and attentive processes. Therefore, the operation of the IDIAP BCI provides strong evidence, if any, that users possess the capability to endorse their own desires and to access interpersonal relationships. Briefly, the operation of this system provides strong evidence that users meet a crucial precondition for autonomy [17] in both individual and relational action spaces.

There is an additional inferential step that the observation of BCI operation does not allow one to undertake with the same confidence. The conclusion that the human user of a BCI system is willing to execute *some* action from the repertoire of actions contemplated by the communication protocol does not license the conclusion that the observed machine behaviour coincides with a reflectively endorsed action. One may appeal to sources of user intent misclassification reviewed in the previous section to question the identification between intended and performed action. Similar doubts about the agreement of user intents with observed behaviours are amplified in communication events involving a *sequence* of user intent interpretations.

Let us turn to consider again question *a* in the light of these observations about the agreement between user intents and actual BCI-controlled behaviours: "What is the binding value of informed consent and last will that locked-in patients express by means of a BCI?" One may appeal to the epistemic uncertainty affecting BCI classifications of user intents in order to question the binding value of those BCI-enabled statements. After all, a misclassification of neural signals which convey a few bits of information may utterly change the meaning of sentences composed by means of a BCI-operated virtual keyboard. In his novel *The History of the Siege of Lisbon*, José Saramago imagines proof-reader Raimundo Silva altering the canonical account of the siege of Lisbon by adding a 'not' in the page proofs of a history book: on their way to Jerusalem the Crusaders did *not* help the king of Portugal against the Moors. Thus, the right to exercise autonomous action by those who are affected by severe motor disabilities appears to be insufficiently protected by current BCI communication protocols, at least insofar as internet transactions, informed consent, and last will statements are concerned. Are there viable technical solutions to improve these communication protocols and protect more effectively the right of disabled people to participate in social life?[10] It is not clear that error detection and correction procedures involving, say, repetition schemes [18] can be effectively applied in order to defuse doubts about the agreement of BCI-rendered statements and user intents. This difficulty derives from the fact that in brain-to-computer communication the crucial channel encoding of classical information theory is completely internal to the message transmitter—namely, internal to human users involved in BCI communication processes. Moreover, the viability of error-detection and correction approaches must be carefully evaluated on account of the restricted channel capacity of current non-invasive BCIs and the additional fatigue that repetition imposes on users.

Shared Control of Robotic Systems and Liability Issues

Let us now turn to explore the ethical dimension of the shared control issue when the BCI-controlled peripheral device is a robotic system. The IDIAP BCI enables one to control a robotic wheelchair navigating through the rooms and corridors of some building floor [6]. For this purpose, a special communication protocol was introduced to control six high-level robotic behaviours (forward, stop, left turn, right turn, left wall following, and right wall following). Two additional low-level behaviours—"obstacle avoidance" and "smooth turning"—are governed by a behaviour-based robotic controller [19], without requiring any

[10] The *Charter of Fundamental Rights of the European Union*, art. 26, states: "The Union recognizes and respects the right of persons with disabilities to benefit from measures designed to ensure their independence, social and occupational integration and participation in the life of the community".

sort of user intervention. Each transition between high-level behaviours is triggered by a combination of two elements: one user mental task (taken from some suitable triad of mental tasks) and one of the six perceptual states (free space, left or right wall, left, right or front obstacle) entered by the robot on the basis of its sensory readings. Clearly, this combinatorics of user mental tasks and robot perceptual states results into a richer repertoire of user commands: "Briefly, the interpretation of a mental state depends on the perceptual state of the robot. Thus, in an open space the mental state #2 means "left turn" while the same mental state is interpreted as "follow left wall" if a wall is detected on the left-hand side." [15, p. 249].

In this brain-to-computer communication context, interpretive errors chiefly arise from two different sources. In addition to direct BCI misclassifications of user mental states, a discrepancy may arise in the semantic contexts identified by human user and robotic system, respectively. More specifically, the perceptual state that human users attribute to the robot—after observing and evaluating the relationship between the robot and its environment—may not coincide with the actual perceptual state entered by the robot. Indeed, limitations of the robot's sensory apparatus and noise affecting its sensory readings may cause the robot to enter a perceptual state which does not correspond to the actual situation (e. g., the robot may misinterpret a "front obstacle" situation for a "free space" situation on this account, propagating this perceptual categorization error onto a user intent categorization error). Similarly, one cannot altogether exclude user misperceptions of robot-environment relationships—on account of, say, poor illumination conditions—eventually resulting in the rehearsal of inappropriate mental tasks. Discrepancies of this sort give rise to a serious control problem, insofar as the robot's perceptual state summarizes all relevant contextual information for user intent identification.

One should be careful to note that a discrepancy between user intents and actual robotic navigation trajectory may arise even when user intents are correctly identified, insofar as the trajectory of a mobile robot depends on its interactions with the environment. In particular, these discrepancies may arise when robot navigation is sensitive to small perturbations of initial conditions or sensor noise piling up in series of sensory readings [20]. In general, for a user motor intent to be fulfilled, the following conditions must be jointly satisfied: (i) correct classification of the mental task that the user is currently performing; (ii) identity of user and robot perceptual states; (iii) correct execution of low-level and high-level behaviours by the behaviour-based controller, and (iv) absence of sufficiently large perturbations affecting the robot's sensory and motor apparatuses.

Let us now turn to consider question *b* in the light of the above epistemological reflections: "Who is responsible for damages caused by a brain-actuated mobile robot?" The overall epistemic context of this responsibility ascription problem is such that users, programmers, and manufacturers are not invariably in the position to predict exactly and certify what this system will actually do in its intended operational environment. This epistemic predicament results from a variety of converging factors, which prominently include our limited understanding and control of the behaviour of learned rules, instability in neural correlates of mental tasks, uncertainties affecting robot sensory readings, and environmental sources of trajectory perturbation. Accordingly, there are conceivable circumstances in which programmers, manufacturers, and users alike—having taken proper care of each design, implementation, and operational aspect—are not morally blameworthy for the damaging event caused by a brain-actuated mobile robot. And if no one is blameworthy, then question *b* must be addressed within an *objective* rather than a *moral* responsibility conceptual framework.

A variety of conceptual and technical tools have been put in place, during the historical development of legal doctrines and systems, to solve objective responsibility or liability ascription problems arising from the use of machines. One may doubt that that these conceptual and technical tools enable one to cope with liability problems concerning robots in general, and brain-actuated robots in particular, in view of the fact that these innovative machines can learn from their experience, manifesting adaptive behaviours that are occasionally difficult to predict and control [21]. One should be careful to note, however, that issues of learning, adaptation, and limited predictability typically arise in liability problems concerning the behaviour of humans and other biological systems. Thus, in developing suitable liability policies for brain-actuated robots, one can rely on a large body of legal knowledge, regulations, and casuistry concerning liability problems arising in connection with the behaviour of these biological systems. Consider, from this perspective, the inability

of BCI users to predict exactly and control the behaviour of brain-actuated robots. Epistemically, this inability is meaningfully related to the inability of employers to predict exactly and control the behaviour of their employees. Employers incurring in this epistemic predicament are nevertheless held to be vicariously liable for many kinds of actions performed by their subordinates. For example, if a damaging event occurs as a consequence of actions performed by employees who are pursuing the interest of their employer, then this employer is usually regarded as liable under the legal doctrine summarized in the Latin formula *Qui facit per alium facit per se*. Furthermore, in view of the fact that training and learning are crucially involved in BCI contexts, the inability of BCI users to predict and control brain-actuated devices is meaningfully related to the inability of dog owners to curb properly trained dogs in every possible circumstance, and even to the inability of parents to predict and control the behaviour of their children. Parents are nevertheless held to be vicariously liable for many kinds of damaging events caused by their children, and pet owners are liable for damages caused by their pets. Finally, let us note that producers of goods can be held liable for damaging events that are difficult to predict and control, on the basis of economic considerations aptly summarized in the Latin formula *ubi commoda ibi incommoda:* expected producer profit provides there a basis for ascribing liability and determining compensations to consumers and society at large.[11] On the whole, these broad analogies suggest that relatively small adjustments, if any, of extant ethical and legal frameworks may be needed to cope with liability issues arising in connection with brain-actuated robots.

Non-invasive BCIs and Personhood

More challenging conceptual and policy issues arise in connection with the pro-active protection of personal identity by means of BCI therapies, and with sub-personal uses of human beings in BCI-enabled operation environments. Let us examine both issues in turn.

BCI Therapies and Personal Identity Persistence

The problem of explaining the inability to learn and operate BCI systems by subjects in the completely locked-in state (CLIS) suggest, in my view, a novel ethical perspective on BCIs as therapies enabling one to protect personal identity.

Niels Birbaumer pointed out that among ALS patients trained with a non-invasive BCI, none of those who were trained *after* entering a CLIS were able to acquire stable communication abilities [3, p. 524]. Birbaumer advanced two competing explanations for this observation. According to the first explanation, the onset of CLIS is accompanied by a generalized decline of perception, thinking, and attention abilities, which prevents CLIS patients from learning to use a BCI. The second explanation hinges on the hypothesis that the development and sustained preservation of purposive thinking crucially involves a reinforcement stage, concerning the verification of intended consequences of actions. This hypothesis is advanced in the framework of so-called motor theories of thinking, according to which thinking develops as a means for—and is sustained by—effective animal motion: "As early as the 19th century, the 'motor theory of thinking' hypothesized that thinking and imagery cannot be sustained if the contingency between an intention and its external consequence is completely interrupted for a long time period." [12, p. 481]. Moreover, this hypothesis is consistent with models of motor planning that crucially involve a comparison between expected and actual consequences of motor plan execution [22, 23]. The reinforcement stage required by motor theories of thinking is hardly ever accessed in a CLIS subject. The sequence intention-action-consequence-verification cannot be enacted autonomously; it is occasionally completed through the intermediary of caretakers who happen to fulfil the patient's current desire. Therefore, thinking and imagery are no longer sustained, and the related ability to learn and operate a BCI fades away in a CLIS patient.

If the first explanation is correct, then learning how to use a BCI does not contrast a progressive and

[11] It is not clear, however, that the more appropriate liability ascription policies for brain-actuated robots in the near future will be those based on economically oriented criteria, in view of the free exchange of technological resources which has become standard practice within the BCI research community: "The non-invasive BCI community overcame commercial temptations with the BCI 2000 website allowing laboratories worldwide access to the necessary hard- and software." [12, p. 482].

generalized decline of perception, thinking, and attention. If the second explanation is correct, however, then learning how to use a BCI before the onset of CLIS in ALS patients may prevent the extinction of thinking and imagery, insofar as the sequence intention-action-consequence-verification is preserved through BCI operation. Accordingly, one should teach BCI operation to ALS patients, in order to preserve their thinking abilities, and ultimately to preserve their status of persons, to the extent that goal-oriented thinking is a central feature of a person. This conditional conclusion provides a distinctive ethical motivation for testing the hypotheses under which the second explanation holds. If the hypotheses leading to the second explanation survive severe empirical tests, then the additional ethical issue must be addressed of shaping fair access policies for preventive medical interventions, taking into account different individual proficiency in BCI use [24] and limited medical resources.[12]

Sub-Personal Mental Agencies in BCI Interactions

The use I will make here of the term 'sub-personal' is grounded into the distinction—originally introduced in the philosophy of mind—between personal level and sub-personal level explanations of human behaviours and mental processes [25, 26]. To illustrate the distinction, consider mental states and processes involved in mental arithmetic tasks, such as iterated mental subtraction. In explaining why and how these mental tasks are performed, one can hardly dispense with the intentional vocabulary of personal level explanations—insofar as one voluntarily initiates and intentionally iterates this calculation procedure. In contrast with this, consider the mental states and processes involved in generating the Kanizsa triangle visual illusion [27]. These mental states and processes are neither initiated nor acted on intentionally. In order to explain why and how these mental states and processes are produced, one usually invokes the deployment of sub-personal perceptual and cognitive abilities that are impenetrable to voluntary activation, intentional modulation, and conscious reflection.

Let us now explore the distinction between personal and sub-personal explanations in connection with the (neural correlates of) mental states and processes identified by BCI systems. The IDIAP BCI looks for neural correlates of mental arithmetic, mental rotation, and motor imagination processes that are voluntarily initiated and consciously controlled by enacting subjects. In contrast with this, the Co3 (Cortically Coupled Computer) Vision System, which crucially involves an EEG-based BCI, identifies neural correlates of mental states that are unconscious and impenetrable to intentional modulation. The Co3 system identifies neural correlates of perceptual classification cues in order to search for special kinds of target images into large image databases. Cooperative human-computer problem solving systems of this kind pave the way to what one may appropriately call a sub-personal use of human beings in HCI, insofar as the role of human beings involved in this cooperative task is accounted without appealing to any of the more characteristic features of human mentality and personhood. Let's see.

Consider the task of distinguishing images which contain a human face from other sorts of images. An early face-selective EEG response was recently isolated, which is consistent with the N170 negative ERP—a face recognition cue arising 170 ms after the visual presentation of a face. This empirical finding furnishes the empirical basis for developing EEG feature classification rules that are predictive of a face/non-face categorization outcome by means of a supervised machine learning procedure [28]. In their turn, these classification rules open up the technological possibility of overcoming limitations of state-of-art computer vision systems, insofar as these systems are a long way from matching human performance in visual classification tasks as far as accuracy and rapidity are concerned. These neuroscientific findings and ICT technological developments converge in the design of the Co3 Vision System, which identifies neural correlates of perceptual classification cues in order to sort out target images from large image databases. In experimental trials, image sequences containing randomly positioned target images were shown to 5 human subjects using a rapid serial presentation protocol. The reported detection average

[12] This issue is examined in the light of the distinction between negative and positive rights in [11]. Moreover, Fenton and Alpert discuss possible enhancement effects in LIS subjects deriving from the use of BCI systems in the light of extended mind theories in the philosophy of mind, according to which BCI-controlled peripheral devices may enable one to augment neural structures for cognitive processing [11, p. 127].

accuracy is 92% relative to the repositioning of target images to the top 10% of the image stack from their random position in a sequence of 2500 images [9].

Prospective applications of similar systems include cooperative human-machine classification of large amounts of video and still images for security, surveillance, and military purposes. Notably, the U.S. Defence Advanced Research Projects Office (DARPA) is supporting the NIA (Neurotechnology for Intelligence Analysts) Program. This program "seeks to identify robust brain signals that are amenable to recording in an operational environment and process these in real time to select images worthy of further review."[13]

Electrophysiological and ICT technologies enabling military, security, and surveillance applications of cooperative perceptual classification systems essentially coincide with the BCI technologies enabling the control of virtual keyboards or robotic wheelchairs. From a cognitive perspective, however, there is a substantive difference between the resulting BCI systems. In the IDIAP BCI, signal analysis and translation modules are ultimately geared so as to identify and fulfil specific motor intents, through the intermediary of voluntarily initiated and consciously controlled mental processes. There is no such connection between human user intents and Co3 Vision System operations. Human subjects who are fully competent about the functional role played by their unconscious perceptual processing in this system are not thereby put in the position to select and modulate the sorting operations which will be actually performed. Therefore, the control transfer required by BCI brain-to-computer communication protocols appears to be particularly extensive in the Co3 Vision System, insofar as BCI users do not exert any kind of intentional control on machine behaviours, let alone their piecemeal execution. Differences in delegation contents emerging from comparative analyses of systems such as the IDIAP BCI and Co3 Vision are likely to emerge in connection with different ways to achieve the same BCI functionality as well. In these cases, design choices should be made by scientists with the aim of minimizing the amount of control transfer which is needed to achieve intended BCI functionality. Accordingly, informed consent requests should come with a description of what is being transferred, and an explanation of how the control function was designed with the aim of requiring a minimal amount of control transfer from human users to machine.

The Co3 Vision system provides a vivid illustration of the possibility of plugging into a BCI system, qua sub-personal cognitive agency, someone who is unaware of the purposes one is using her mental processing for. The sub-personal cognitive agency employed in the Co3 Vision System performs categorization of visually presented objects. An additional possibility of making a sub-personal use of human beings in the framework of non-invasive BCI systems is suggested by mental imagery experiments performed on subjects who were found to fulfil the criteria for a diagnosis of vegetative state [29]. Some of these patients were given spoken instructions to perform mental imagery tasks that are in principle suitable for operating a BCI such as the IDIAP BCI. It turns out that the neural responses of these patients to the verbal instruction of imagining a game of tennis, as recorded by fMRI scanning of the supplementary motor area, are indistinguishable from those observed in healthy volunteers performing the same imagery task. Thus, from the narrow perspective of peripheral device activation induced by a mental imagery task, one of those healthy volunteers and one of those verbally instructed patients appear to be interchangeable cognitive agencies.[14]

Let us now turn to examine question *c*: "Is human dignity jeopardized by the use of unconscious or preconscious brain information processing in BCI-enabled, human-machine cooperative problem solving?" Clearly, obtaining informed consent and making people aware of their functional role as sub-personal cognitive agencies does not suffice to defuse worries about the protection of human dignity in the context of a BCI-enabled cooperative problem solving system. Indeed, consider from this perspective the scenario of a company which addresses the widespread need for more efficient searches into huge

[13] http://www.darpa.mil/dso/thrusts/trainhu/nia/index.htm (site visited on February 12, 2009).

[14] The authors of this study go as far as claiming that "...the presence of reproducible and task-dependent responses to command without the need for any practice or training suggests a method by which some non-communicative patients, including those diagnosed as vegetative,... may be able to use their residual cognitive capabilities to communicate their thoughts to those around them by modulating their own neural activity." [29] p. 1402.

digitalized archives of photographs and video streams. If this company decides to operate BCI-enabled cooperative problem solving systems, then job openings will be advertised for unskilled workers willing to sell their own unconscious perceptual processing abilities. It is a commonplace that the repetitive and "mechanical" activity of an assembly line worker usually engages very little conscious reflection and attention levels. Clearly, the possibility of exercising human thinking would be even more extensively curtailed in this envisaged BCI context. Moreover, these workers will be actively discouraged from "thinking on the side" during working shifts, insofar as this activity is a main source of perturbation which may disrupt the system proper functioning.

This envisaged scenario suggests the opportunity of scrutinizing cooperative perceptual problem solving systems that will step outside of research labs for violations of Kant's formula of humanity—according to which we should never act in such a way as to treat humanity as a mere means to our ends, but always as an end in itself [30]. More important, I surmise that a novel form of alienation is lurking here. If intentional and conscious mental processing is to be counted as a typically human feature, then the unskilled workers that are plugged into these BCI-enabled cooperative problem solving systems are deprived of a typically human feature, and made impersonal to the extent that their function is entirely dependent on their sub-personal cognitive abilities. This novel form of alienation is closely related to—and arguably subsumed by—Marx's notion of alienation from one's own species, insofar as intentional and conscious thinking is a typical trait of *homo sapiens*.

Concluding Remark: BCI Visions and Personhood

In these final observations, I will waive the triaging constraint applied in previous sections, that is, the precept of narrowing down ethical reflection to imminent developments of BCI technologies and systems. In fact, I will briefly comment on the import of BCI visions for the narrative dimension of personal identity,[15] taking a leap beyond present and imminent developments of neurofeedback systems that are based on BCI technologies.

Prototypes of BCI neurofeedback systems currently include applications for preventing epileptic seizures and treating hyperactive children. When provided with EEG-detected information about their own states of neural excitability, epileptic subjects learned to control their brain activity so as to achieve a significant reduction of seizure onset. Hyperactivity reduction therapies for children affected by ADHD (Attention-Deficit/Hyperactivity Disorder) take advantage of similar BCI neurofeedback systems, in addition to BCI-controlled video-games which reward user relaxation efforts. Alertness detectors based on similar BCI technologies are expected to lead to improved safety and efficiency in the workplace, insofar as a critical decline in alertness levels can be detected in intensive workflow situations, such as those involving security personnel, airplane pilots, and truck drivers.

These various BCI systems must be deployed in accordance with general ethical guidelines concerning access to sensitive data of patients and workers, notably for the purpose of preventing discriminations in the workplace (cp. question *d* above). Moreover, BCI-controlled games and training systems in simulation environments will require extensive risk analysis, especially concerning the effect on the plastic brains of young people. An appeal to the precautionary principle appears to be fitting in this case, if precaution is sensibly construed as a warning about the relative deficiency of deep models and experimentally supported scientific knowledge about brain plasticity (cp. question *e* above).[16]

José del R. Millán envisages a scenario of brain-resonating technologies based on future developments of BCI technologies for neurofeedback and brain state detecting systems [31]. These systems will enable one to extend and refine knowledge about one's own brain and mental states by immersing the brain into a BCI-enabled network of ICT and robotic agents. Consider, for example, the confluence of BCI and emotional computing technologies. In emotional computing

[15] See, for discussion, [39] pp. 249-271 and, more specifically in connection with BCI systems, [40].

[16] Rather than as a paralyzing maxim which uniformly blocks the use of technologies if one cannot exclude undesirable consequences with absolute scientific certainty. This construal of the precautionary principle is arguably incoherent, insofar as it is oblivious to the fact that scientific theories and models are inherently fallible.

systems, biometric data are used as input to a knowledge-based system which advances and revises hypotheses about the current emotional states of their users. By merging these data—which include skin conductance, muscular contraction, posture, heart rate, blood pressure, voice intonation, face expressions, and so on[17]—with EEG data about neural correlates of mental states, these systems might be able to provide human users with richer and better guesses about their own mental states, thereby facilitating self-knowledge and enhancing human capabilities for introspection and reflection. These forms of HCI involving BCI technologies might have a profound impact on user personality and personal identity, insofar as these interactions may deeply affect the user's self-interpretations, self-concept, and narrative styles about self-related beliefs. Briefly, these BCI systems might become novel cybernetic companions helping human beings to explore uncharted psychological territories and to pursue in unprecedented ways the know thyself precept.[18]

Acknowledgments I wish to thank an anonymous reviewer, Giuseppe Trautteur, Federica Lucivero, and Giovanni Boniolo for helpful and stimulating comments. I benefited from discussions on BCI systems and ethics with Febo Cincotti, Edoardo Datteri, José del R. Millán, Donatella Mattia, Stefano Rodotà, and Matteo Santoro.

References

1. Birbaumer, N., N. Ghanayim, T. Hinterberger, B. Kotchoubey, A. Kuebler, J. Perelmouter, E. Taub, and H. Flor. 1999. A spelling device for the paralyzed. *Nature* 398:297–298.
2. Hochberg, L.R., M.D. Serruya, G.M. Friehs, J.A. Mukand, M. Saleh, A.H. Caplan, A. Branner, D. Chen, R.D. Penn, and J.P. Donoghue. 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* 442:164–171.
3. Birbaumer, N. 2006a. Breaking the silence: Brain-computer interfaces for communication and motor control. *Psychophysiology* 43:517–532.
4. Wolpaw, J.R., N. Birbaumer, D.J. McFarland, G. Purtscheller, and T.M. Vaughan. 2002. Brain-computer interfaces for communication and control. *Clinical Neurophysiology* 113:767–791.
5. Millán, J. del R., F. Renkens, J. Mouriño, and W. Gerstner. 2004. Brain-actuated interaction. *Artificial Intelligence* 159:241–259.
6. Galán, F., M. Nuttin, E. Lew, P.W. Ferrez, G. Vanacker, J. Philips, and J. del. R. Millán. 2008. A brain-actuated wheelchair: Asynchronous and non-invasive brain-computer interfaces for continuous control of robots. *Clinical Neurophysiology* 119:2159–2169.
7. Friedman, D., R. Leeb, L. Dikovsky, M. Reiner, G. Pfurtscheller, and M. Slater. 2007. Controlling a virtual body by thought in a highly immersive virtual environment, in *GRAPP 2007*, Barcelona, Spain, 83–90.
8. Nijholt, A., D. Tan, A. Brendan, J. del R. Millán, B. Graimann. 2008. Brain-computer interfaces for HCI and games, in *Proceedings of CHI'08*, ACM, pp. 3225–3228.
9. Gerson, A.D., L.C. Parra, and P. Sajda. 2006. Cortically coupled computer vision for rapid image search. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 14(2):174–179.
10. Yahud, S., and N.A. Abu Osman. 2007. Prosthetic hand for the brain-computer interface system. *IFMBE Proceedings* 15:643–646. Springer, Berlin.
11. Fenton, A., and S. Alpert. 2008. Extending our view on using BCIs for locked-in syndrome. *Neuroethics* 1:119–132.
12. Birbaumer, N. 2006b. Brain-computer interface research: Coming of age. *Clinical Neurophysiology* 117:479–483.
13. Buxton, R.B. 2002 *An introduction to functional magnetic resonance imaging: Principles and techniques*. Cambridge UP.
14. Linderman, M. D., G. Santhanam, C.T. Kemere, V. Gilja, S. O'Driscoll, B.M. Yu, A. Afshar, S.I. Ryu, K.V. Shenoy, T. H. Meng. 2008. Signal Processing Challenges for Neural Prosthetes. A Review of State-of-Art Systems, *IEEE Signal Processing Magazine* 18.
15. Millán, J. del R. 2004. On the Need for On-line Learning in Brain-Computer Interfaces. *International Joint Conference on Neural Networks*.
16. Vapnik, V. 2000. *The nature of statistical learning theory*. 2nd ed. New York: Springer.
17. Reath, A. 1999. Autonomy, ethical. In *Routledge encyclopedia of philosophy*, ed. E. Craig. London: Routledge.
18. MacKay, D. 2003. *Information theory, inference, and learning algorithms*. Cambridge UP.
19. Arkin, R. 1998. *Behavior-based robotics*. Cambridge: MIT.
20. Nehmzow, U. 2006. *Scientific methods in mobile robotics*. London: Springer.
21. Matthias, A. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6:175–183.
22. Miall, R.C., and D.M. Wolpert. 1996. Forward models for physiological motor control. *Neural Networks* 9:1265–1279.
23. Kawato, M. 1999. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology* 9:718–727.

[17] Early ethical reflection on affective computing are found in [41].

[18] This scenario reminds one of the psychoanalytic variation of the know thyself maxim that Freud set out as a main goal for psychoanalytic interactions, that is, "…to strengthen the ego, to make it more independent of the superego, to widen its field of perception and enlarge its organization, so that it can appropriate fresh portions of the id. Where id was, there ego shall be. It is a work of culture - not unlike the draining of the Zuider Zee ." [42 p. 80 of the English translation].

24. Bufalari, S., F. Cincotti, F. Babiloni, L. Giuliani, M.G. Marciani, and D. Mattia. 2007. EEG patterns during motor imagery based volitional control of a brain computer interface. *International Journal of Electromagnetism* 9:214–219.
25. Dennett, D. 1969. *Content and consciousness*. London: Routledge & Kegan Paul.
26. Hornsby, J. 2000. Personal and Sub-Personal: A Defence of Dennett's Original Distinction. In *New Essays on Psychological Explanation*, Special Issue of *Philosophical Explorations*, eds. M. Elton, and J. Bermudez, 6–24.
27. Kanizsa, G. 1955. Margini quasi-percettivi in campi con stimolazione omogenea. *Rivista di Psicologia* 49:7–30.
28. Philiastides, M.G., and P. Sajda. 2006. Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex* 16:509–518.
29. Owen, A.M., M.R. Coleman, M. Boly, M.H. Davis, S. Laureys, and J.D. Pickard. 2006. Detecting awareness in the vegetative state. *Science* 313:1402.
30. Kant, I. 1983. Grounding for the Metaphysics of Morals. In *Kant's Ethical Philosophy*, ed. J.W. Ellington. Indianapolis: Hackett.
31. Millán, J. del R. 2007. Tapping the mind or resonating minds? In *European visions for the knowledge age, a quest for new horizon in the information society*, ed. P.T. Kidd, 125–132. Macclesfield: Cheshire Henbury.
32. Farah, M.J. 2002. Emerging ethical issues in neuroscience. *Nature Neuroscience* 5:1123–1129.
33. Nordmann, A. 2007. If and then: A critique of speculative nanoethics. *Nanoethics* 1:31–46.
34. Warwick, K. 2003. Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology* 5:131–137.
35. Tamburrini, G. 2006. Artificial intelligence and Popper's solution to the problem of induction. In *Karl Popper: A centenary assessment. Metaphysics and epistemology*, vol. 2, eds. I. Jarvie, K. Milford, and D. Miller, 265–284. London: Ashgate.
36. Santoro, M., D. Marino, and G. Tamburrini. 2008. Robots interacting with humans. From epistemic risk to responsibility. *Artificial Intelligence and Society* 22:301–314.
37. Christman, J. 2003. Autonomy in moral and political philosophy. *Stanford encyclopedia of philosophy*, http://plato.stanford.edu/entries/autonomy-moral/
38. Hansson, S.O. 2007. The ethics of enabling technology. *Cambridge Quarterly of Healthcare Ethics* 16:257–267.
39. Merkel, R., G. Boer, J. Fegert, T. Galert, D. Hartmann, B. Nuttin, and S. Rosahl. 2007. *Intervening in the Brain. Changing psyche and society*. Berlin: Springer.
40. Lucivero, F., and G. Tamburrini. 2008. Ethical monitoring of brain-machine interfaces, A note on personal identity and autonomy. *AI and Society* 22:449–460.
41. Reynolds, C., and R.W.Picard. 2004. Affective sensors, privacy, and ethical contracts. *Proceedings of CHI'04*, ACM, 1103–1106.
42. Freud, S. 1933. New introductory lectures on psychoanalysis. *The standard edition of the complete psychological works of Sigmund Freud*, vol. 22, 1–182. London: Hoghart.