Motor Control Brain Implants: Design Considerations

DT 313 : Technology Ethics and AI
Ajay Ramesh - IMT2017502
Anirudh C - IMT2017006
Kocherla Nithin Raj - IMT2017511
Ronak Doshi - IMT2017523
November 30, 2020

Abstract

The technology of brain implants is attracting growing research attention due to its immense potential in medical applications. Brain implants can act as a substitute for dysfunctional parts of the brain and help humans regain lost perceptive and physical abilities, e.g. biomedical prosthesis and sensory substitution. The use of brain implants however, has a wide range of ethical and philosophical implications ranging from privacy and security to free will and moral responsibility. In this work, we examine such concerns that arise from the technology's general base design and application by drawing upon the ideas of humanist ethics, feminist ethics, moral responsibility, care ethics, free will and human autonomy. Additionally, we address these concerns by proposing modifications in each of the base design elements.

Introduction

A brain computer interface (BCI) serves as a pathway between the brain and an external device or component (Krucoff et al., 2016). A brain implant is a kind of brain computer interface which aims to physically substitute for a part of the human brain. These are particularly useful for people who have lost certain motor abilities. A brain implant could facilitate motor actions which help disabled people perform daily activities, restore a sense of normalcy in their lives and integrate into society. However, this integration is non-trivial and requires careful consideration.

Technological advancements in BCIs over the last decade have been constantly accompanied by discourse in their ethical and philosophical concerns. These concerns arise from the different elements of BCI system design and include privacy, security, autonomy, personhood, stigma, self-determination, responsibility, and consent (Wolkenstein et al., 2018). The usage of BCIs and the integration of BCI users into society reveal new manifestations of these ethical concerns and require urgent resolution in order to maintain societal harmony despite the potentially ubiquitous nature of this technology. In this article, we examine the ethical implications of BCI use, the role of current technical design in giving rise to these ethical concerns, and finally propose modifications to the general BCI framework to mitigate certain concerns.

Design

A high-level generic design of a BCI system is shown in Fig. 1. The implant collects raw information in the form of brain signals which are pre-processed to extract relevant features that the AI requires to make the final decision. The AI's decision is translated back into a brain signal to stimulate the part of the brain that executes motor actions in a given environment via the muscles, limbs etc. The user perceives and comprehends his/her actions in the

environment, which in turn can influence subsequent actions. This user perception accompanied with the information delivered by the system to the user regarding the result of the AI's decision forms the feedback element.

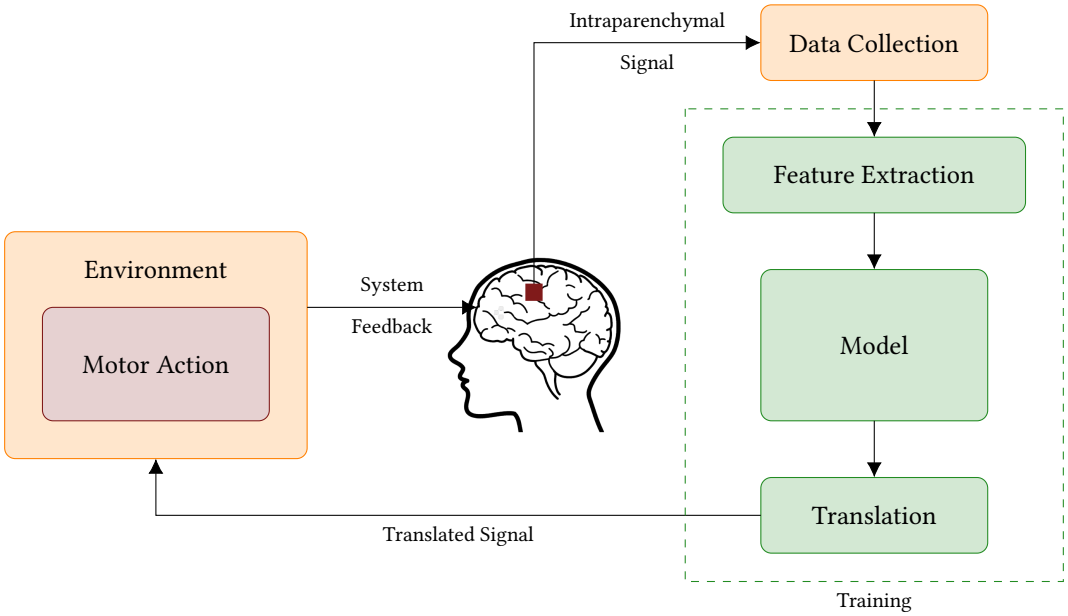

Figure 1: Rudimentary Design Structure of Implant

Data and Data-collection

Artificial Intelligence systems are driven by large amounts of data which define algorithmic behavior. Thus, to design an AI-powered brain implant, it is important to analyze and address ethical concerns in the data and its collection methods.

Brain-computer interfaces (BCI) usually collect EEG data by non-invasively placing electrodes on the skull. However, these EEG signals are noisy and the resulting performance is sub-optimal. A brain implant does not just interpret brain signals but also responds with an appropriate stimulus (bi-directional). Therefore, electrodes need to be implanted invasively at the cortex of the brain to collect *Intraparenchymal signals* (NeuroTechEDU, n.d.) which are more efficient and informative compared to EEG signals captured via non-invasive techniques. Hence, building a bi-directional BCI that controls motor action requires invasive techniques to collect data and as a result, ethical concerns regarding consent and human rights will arise.

The risks and complications associated with surgical implantation include hemorrhage (1.3–4%), infection (2.8–6.1%), lead migration, misplacement or breakage (5.1%), and even death (0.4%) (Clausen, 2010). These issues pertain to the domain of medical ethics and will not be addressed in this article. Although consent of the subject can be obtained before the procedure, the potential impact of using the brain implant in society demands proper consideration.

Privacy and security are of major concern while collecting data. The implant can collect data 24/7, which raises the question of what data should be collected and when. Collected Intraparenchymal signals can monitor a patient's emotional and cognitive reactions. These metrics allow operators to obtain the subject's feelings or thoughts about sensitive topics such as sexuality, politics, and autonomy. In essence, they provide deep insight into a subject's

preferences, which may not be welcome. Every human, as an autonomous being, requires respect for personal privacy. However, we must first consider why maintaining privacy is crucial. Drawing from Humanist ethics (Dierksmeier, 2011), to lose control of one's personal information is in some measure to lose control of one's life and one's dignity. Therefore, even if privacy is not in itself a fundamental right, it is necessary to protect other fundamental rights such as *Misuse of Personal Information* , *Privacy and Relationship*, *Autonomy* and *Human Dignity*. Hence the BCI raises security and privacy concerns which can result in the subject losing the ability to control his/her own data. A possible design solution to address these concerns would be to encrypt the collected data and apply privacy-preserving machine learning techniques on the encrypted data (Liu et al., 2020). This will help in tackling the security and privacy issues, but not the issue of sovereignty of the subject on its own data. Hence, giving the subject control over the data collection process of the implant will help in addressing the issue.

Privacy concerns aside, biases in the data also pose a major ethical concern. The BCI system helps to emulate motor actions that technically involve executing a sequence of movement states which reflect the body language of the subject. Body language can convey a wide range of meanings and is a characteristic that varies from group to group (Ren & Zhi-peng, 2014; Kendon, 2017; Feyereisen, 1994). For example, body language for a particular action may vary for males and females. Data majorly collected from only one particular gender may result in a dataset biased towards a particular gender. Therefore the collection of data must be equally distributed across different people and cultures in order to avoid bias. There is also the need to maintain transparency in the dataset. Most of the available BCI datasets (BNCI Horizon 2020, n.d.) contain raw data with only the patient ID and are not encrypted, thereby exposing private information. Moreover, they do not reveal any kind of demographics, as a result of which biases remain hidden. To sum up the necessary requirements, it is important to bring in a level of transparency in the dataset by disclosing the overall demographics while at the same time hiding individual demography.

Model Engineering

The underlying algorithm that accomplishes the required functionality of the brain implant gives rise to issues of its own; both in the design stage and in the deployment stage. Such an algorithm/model can be viewed as translating an individual's *intention* to perform an action to the required action (Stephen Rainey, 2020). Existing applications (Bockbrader, 2019) aiming to achieve similar goals as the brain implant, represent the components of the model as

- neural interfaces for recording/stimulating the brain;
- signal processing; and
- decoders and actuators for inducing the command to perform the motor action.

The decomposition of the system into these components is fairly restrictive as it corresponds to a very specific application. The novelty of the proposed implant entails that a technical design cannot be ascertained at this point in time. Hence, a generic set of components can be considered as depicted in Fig. 1.

Design Goals

The mental and physical processes of the human body are governed by the so called *affective states* (Shan, 2012; Noroozi et al., 2018). These are *psycho-physiological constructs* that pertain

to the experience of *feeling, emotion or mood* and their manifestation in physical actions. The underlying component with respect to motor action is that of body language as mentioned in the Data section. The complex state of the mind can give rise to a variety of different body language indicators which other humans implicitly process in any interaction. Hence, the design goals that the model must strive to achieve are two-fold: correctness and representation of body language. Correctness requires the model to induce a motor action that was "intended" by the user accurately, whereas, the representation of body language must build on inferences from the affective states of the mind. These affective states must be captured by the "Feature Extraction" component of the base design in Fig. 1.

Algorithmic Bias

Failure to achieve the aforementioned design goals results in a variety of ethical concerns: unauthentic expression of emotion, inexplicability of certain affect-driven body language, etc. (Steinert & Friedrich, 2020). These concerns can be attributed to the humanist standpoint, that is, individuals have the right and responsibility to shape their lives and specifically, their social perception. The inexplicability that could arise due to a disingenuous reproduction of a user's intended action or body language restricts the user's capability to represent themselves in society. On the other hand, any such algorithm that aims to generalise the subjective and nuanced aspect of social identity has the potential to induce a systemic bias. The bias could result from the data used in a training paradigm for the model, or the underlying decision structure the model uses.

It has been shown that gender is a distinguishing aspect of affective experiences in humans (Moriguchi et al., 2014). Given the existing imbalance in female representation in society, it is very plausible that such a bias against women can arise. Feminist philosophy suggests that the process of generalising any decision-making system gives in to the male-dominated perspective of the world. In the context of body language, the issue is further amplified given that the physical expression of an individual's identity varies across men and women. The social perception of gender coupled with possible inaccuracies of the model can result in a systemic bias against users of such an implant.

Bias arises not just in the gendered context but in a more general social identity context as well and a design solution from the model engineering perspective can be to segment the evaluation criteria of representation of body language into multiple *classes* corresponding to the variety of social identities in a given society. Each class must capture the efficacy and accuracy of the model in representing an individual's identity through their body language. In the context of a machine learning-based model such a metric could be the micro-averaging F1 Score which attempts to quantify the accuracy of a model with respect to each class. In addition to a refinement of the accuracy metrics used to evaluate the model, the test set of patients must be varied as well in order to give the designers a more holistic approach to improving model design and parameters.

The design goals described above provide a general directive to reduce the inherent algorithmic bias, but achieving perfection with respect to model evaluations (in the form of metrics) is not realistic. The general direction of the model design iterations, on the other hand, must of course aim to achieve near-perfect results.

BCI Mediated Actions and Responsibility

The BCI mediated system drives physical actions based on the AI's decisions. However, the AI is not completely prescient and is prone to perform unintended actions or intended actions in an unintended manner. In such cases, it is unclear how moral responsibility must be attributed. Conventional law defines an action as a *willed* movement (Munoz-Conde F., 2006). But do BCI mediated actions qualify as *willed* movement? In order to effectively answer this question and assess moral responsibility, it is important to understand how these situations technically arise and how they are perceived by a third party.

Actions

An unintended action is caused when the AI fails to differentiate between imagination and execution (decision) in the human mind. Many a time we imagine certain actions but do not decide to perform them. For example, in a fit of anger, one might imagine punching someone and yet have no intention of executing the punch. If such an incident involved a person who was not using a BCI, it is straightforward to say that moral responsibility lies with the person who landed the punch, since a normal person has full control over their physical actions. However, it is not as simple if the action was BCI mediated because BCI users do not have full control over their physical actions. Moreover, certain neural activity such as episodes of imagination and memory can be involuntary. The AI must hence be *selective*, i.e, be able to differentiate and discard certain signals and avoid incorrect translation of imagination into action.

A deliberated action could be performed in an unintended manner when the AI fails to understand the relationship between the executive and implementational dimensions of control (J., 2015). Executive control generally remains unchanged and involves setting the end-goal, such as 'getting a glass of water'. The implementational dimension pertains to how this goal can be achieved and is context specific (a glass of water can be fetched in many ways depending on environmental factors). Technically, the AI is designed to execute the implementational dimension. This requires the AI to correctly decode a set of thoughts into goals and actions.

From the above mentioned scenarios, it is clear that a combination of intent, responsibility, consequences, and perception form the definition of a BCI mediated action.

Attributing Responsibility

The attribution of moral responsibility involves assessing the BCI mediated action itself as well as the circumstances in which the action was performed. Firstly, the purpose of use can be considered; whether the use of a BCI was necessary (eg, persons with disabilities) or recreational (Stephen Rainey, 2020). Secondly, whether the consequences were accidental or intentional. It is straightforward to attribute moral responsibility to a user if the action was intentional or a result of recreational use, since the former is a consequence of deliberation and the latter is a case of negligence. However, necessary use with accidental consequences presents a tricky scenario. It can be tackled from two perspectives, viz. the technology itself and the legal framework regarding actions and responsibility.

Targeted modifications in the technology can help avoid certain inadvertent consequences. Human control within this system is imperative in order to include shared awareness of a situation (Dignum, 2017) and ultimately keep the human within the system's loop. Li *et al.* call

this a *human-in-a-loop* system (Li et al., 2014). Re-establishing human control in the system will shift the onus of moral responsibility back to the user/human, and will enable moral evaluation by existing human norms. In an effort to realize this, we propose an additional user feedback in the form of consent. Qualitatively, this involves the AI asking the user to confirm that its decision aligns with the intent of the user. Such a mechanism will greatly reduce the chances of a BCI mediated action being completely unintended. This additional feedback mechanism is explained in greater detail in the next section.

A modification is also required in a legal sense to evaluate a BCI mediated action. As indicated already, the intent and circumstances must be considered before attributing moral responsibility. Although it is very difficult to change the law, 'reasonable adjustments' can be made when evaluating a BCI mediated action (Stephen Rainey, 2020). These adjustments could be similar to those for disabled persons. For instance, (Stephen Rainey, 2020) suggested that if BCI use was necessary, and an action has caused some accidental harm, exceptions can be made since the user could not act at all if not for the device (even though it is error-prone). This reveals the necessity to consider the particular needs and circumstances of a BCI user. Hence, we naturally draw upon the idea of *care ethics* (Dancy, 1992), which takes the concrete needs of particular individuals as the starting point of what must be done (Kaufman-Osborn et al., 2018). Caring involves considering the BCI user's point of view, their objective needs, and what they expect from us (Conn, 1985). As Daniel encouragingly suggests, such a perspective can be integrated as an institutional theory (Engster, 2004).

At the same time, the definition of action needs to be broadened because if it is just 'willed' then every criminal BCI mediated action could claim that the act was not an *action* in the legal sense. Thus, it is important to consider the overall circumstances and situation in order to effectively attribute moral responsibility in the case of both genuinely innocent and criminal actions. In this sense, there is no excuse for harm caused from recreational use because the user would have been fully aware of the possibility of harm from a potentially error-prone device. This would be a serious case of negligence and must be dealt with accordingly.

Feedback Loop

Feedback is defined as the return of information about the result of a process to the system itself. This involves the outputs of a system being routed back as inputs forming a closed loop. In BCI devices such as the proposed AI brain implant, feedback is a quintessential part of the system as it enables the user and the system to adapt to each other. This adaptation is required because the trained model in and of itself can't yield perfect results due to various limitations of the model. Feedback aims at mitigating this error in outcome. The user should be central to this feedback loop. Thus, BCI devices must be thought of as *adaptive closed-loop* systems.

A lot of the present literature mainly utilises the notion of *system feedback* alone as depicted in Fig. 1, i.e., feedback from the system to the user (Kosmyna & Lécuyer, 2017; Mak & Wolpaw, 2009). Such feedback includes displaying the results of an intended action to the user through visual or auditory means by the system and also by the user's perception of his/her environment in which the action takes place, thus helping the individual to modulate their brain signals; essentially, correcting their thought. Significant work has been done in designing paradigms to make such a form of feedback as user friendly as possible. (Kosmyna & Lécuyer, 2017; Mousavi et al., 2017). However, there are ethical issues that arise due to this. Usually, the algorithm is trained offline on data captured from various users. System

feedback tries to make the user adapt to the constructs as dictated by the algorithm of the implant. This implies that users get trained to adapt their thought processes in accordance with the erroneous AI to perform a certain task. Such a phenomenon threatens to erode free will because the user is coerced into acting according to the demands placed by the implant. It also has huge ramifications on the idea of the self because the user's intended action is now an amalgamation of such an action as performed by people in society at large rather than his/her own intent. The latter is mitigated by performing offline training on that particular individual by making the user perform certain intended tasks and recording data (Kosmyna & Lécuyer, 2017). This, however, still is subject to the problem of coercive control exerted by the algorithm in trying to "correct" the individual's intended thoughts. Also, as explained in previous sections, contextual information might become blurred in such a scenario. To summarize, the use of system feedback alone seems to place the agency of human thought more on the BCI system and less on the person using the BCI.

A simple design solution to the ethical concerns raised above would be to add the element of *user feedback*, i.e., feedback from the user to the system, along with system feedback. This would truly close the loop of the system. User feedback would allow not just the user adapting to the system but the other way around as well. Such a form of feedback can either be explicit or implicit (Kosmyna & Lécuyer, 2017). Both of these methods would result in a system where an error signal is fed back to the algorithm which adaptively trains itself. This procedure of active training uses the notion of *reinforcement learning*. A theoretical model based on the idea of a 'reinforcement signal' that incorporates both the notions of explicit and implicit feedback in a common framework has been proposed (Llera et al., 2012). Explicit feedback directly corresponds to the consent mechanism that has been proposed in the previous section. Thus, based on user consent, the BCI actively learns outcomes to the user's thoughts. However, the problem of latency arises here as the consent seeking behaviour of the BCI can tend to increase the time delay between thought and action. However, thanks to the nature of the reinforcement signal, the BCI can also adaptively learn for what sorts of thoughts explicit consent is needed based on user feedback. As a result, the performance of the BCI will improve with time, thereby mitigating the issue of latency associated with explicit consent mechanisms.

This notion of incremental training would aim at a two-way adaptation (system and user) which can mitigate some of the ethical concerns raised earlier. The user now has the flexibility to act in accordance to his/her own thought processes rather than having to be constantly corrected by the system. This would also enable more inclusion in a dignified manner as the user wouldn't be bound to the requirements imposed by the implant. As discussed in the previous section, the addition of a reinforcement signal ensures that the human is central to the feedback loop, shifting the agency, which was completely with the BCI, to the user. In addition, the reinforcement mechanism has the potential to learn the ethical injunctions made by the user in the consent-seeking process, which the BCI in turn replicates later on without asking for explicit consent. The implication of this is that the AI is now transformed, over time, from being subject to total human control to an *artificial moral agent* (Dignum, 2017) (AMA) which derives its agency from the user. Thus, even though the AI has the potential to act with its own agency, the accountability and responsibility is still owned by the user.

Conclusion

In this article, we analyze how ethical issues arise from present BCI systems by relating ethical concerns to normative and natural law frameworks and applying them in the BCI space. Furthermore, in order to address these, modifications have been proposed to the base design (Fig. 1). The proposed design is shown in Fig. 2. The newly proposed elements aim to address the ethical concerns brought about in each of the design steps. The raw information collected by the implant is now encrypted to address privacy concerns. From these signals, along with classical features, features pertaining to affective states are extracted to capture contextual information of an action. Lastly, the addition of the user feedback block to the feedback element of the design enables the human to maintain agency and autonomy, and train the BCI system based on his/her needs.

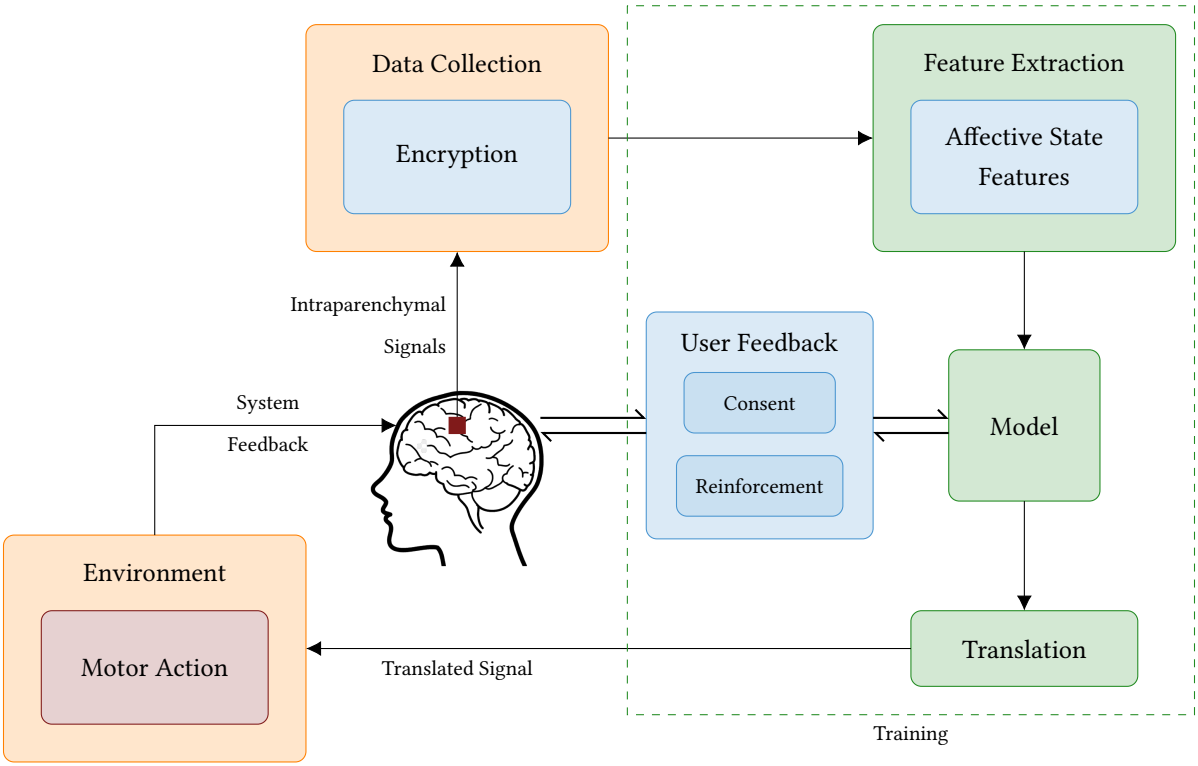

Figure 2: Proposed Design Changes marked in blue

References

Bnci horizon 2020. (n.d.). http://bnci-horizon-2020.eu/database/data-sets. (Accessed: 2020-11-21)

Bockbrader, M. (2019). Upper limb sensorimotor restoration through brain–computer interface technology in tetraparesis. Current Opinion in Biomedical Engineering, 11, 85 - 101. (Biomechanics and Mechanobiology: multiscale modeling ●Novel Biomedical Technologies: Medical devices > point of care (LMIC)) doi: https://doi.org/10.1016/j.cobme.2019.09.002

Clausen, J. (2010, 10). Ethical brain stimulation—neuroethics of deep brain stimulation in research and clinical practice. The European journal of neuroscience, 32, 1152-62. doi: 10.1111/j.1460-9568.2010.07421.x

Conn, W. E. (1985). Caring: A feminine approach to ethics and moral education. by nel noddings. berkeley: University of california press, 1984. 216 pages. Horizons, 12(1), 209–210. doi: 10.1017/S0360966900034824

Dancy, J. (1992). Caring about justice. Philosophy, 67(262), 447–466. doi: 10.1017/S0031819100040651

Dierksmeier, C. (2011, 01). Kant's humanist ethics. In (p. 79-93). doi: 10.1057/9780230314139_6

Dignum, V. (2017). Responsible autonomy. In Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17 (pp. 4698–4704). doi: 10.24963/ijcai.2017/655

Engster, D. (2004). Care ethics and natural law theory: Toward an institutional political theory of caring. Journal of Politics, 66(1), 113-135. doi: https://doi.org/10.1046/j.1468-2508.2004.00144.x

Feyereisen, P. (1994). Hand and mind: What gestures reveal about thought. The American Journal of Psychology, 107(1), 149–155.

J., S. (2015). Conscious control over action. Mind & Language, 30, 320.

Kaufman-Osborn, T., Hirschmann, N. J., Engster, D., Robinson, F., Yeates, N., & Tronto, J. C. (2018). Moral boundaries: A political argument for an ethic of care. by joan tronto. new york: Routledge, 1993. 242 pp. Politics and Gender, 14(4), E18. doi: 10.1017/S1743923X18000417

Kendon, A. (2017, Feb 01). Reflections on the "gesture-first" hypothesis of language origins. Psychonomic Bulletin Review, 24(1), 163-170. doi: 10.3758/s13423-016-1117-3

Kosmyna, N., & Lécuyer, A. (2017, 07). Designing guiding systems for brain-computer interfaces. Frontiers in Human Neuroscience, 11, 396. doi: 10.3389/fnhum.2017.00396

Krucoff, M. O., Rahimpour, S., Slutzky, M., Edgerton, V. R., & Turner, D. (2016). Enhancing nervous system recovery through neurobiologics, neural interface training, and neurorehabilitation. Frontiers in Neuroscience, 10.

Li, W., Sadigh, D., Sastry, S. S., & Seshia, S. A. (2014). Synthesis for human-in-the-loop control systems. In E. Ábrahám & K. Havelund (Eds.), Tools and algorithms for the construction and analysis of systems (pp. 470–484). Berlin, Heidelberg: Springer Berlin Heidelberg.

Liu, Y., Huang, H., Xiao, F., Malekian, R., & Wang, W. (2020, 10). Classification and recognition of encrypted eeg data based on neural network. Journal of Information Security and Applications, 54, 102567. doi: 10.1016/j.jisa.2020.102567

Llera, A., Gómez, V., & Kappen, H. J. (2012). Adaptive classification on brain-computer interfaces using reinforcement signals. Neural Computation, 24(11), 2900-2923. (PMID: 22845827) doi: 10.1162/NECO_a_00348

Mak, J. N., & Wolpaw, J. R. (2009). Clinical applications of brain-computer interfaces: Current state and future prospects. IEEE Reviews in Biomedical Engineering, 2, 187-199. doi: 10.1109/RBME.2009.2035356

Moriguchi, Y., Touroutoglou, A., Dickerson, B. C., & Barrett, L. F. (2014, May). Sex differences in the neural correlates of affective experience. Social cognitive and affective neuroscience, 9(5), 591-600. (23596188[pmid]) doi: 10.1093/scan/nst030

Mousavi, M., Koerner, A. S., Zhang, Q., Noh, E., & de Sa, V. R. (2017). Improving motor imagery bci with user response to feedback. Brain-Computer Interfaces, 4(1-2), 74-86. doi: 10.1080/2326263X.2017.1303253

Munoz-Conde F., C. L. E. (2006). The act requirement as a basic concept of criminal law. Cardozo Law Review, 28, 2461.

Neurotechedu. (n.d.). http://learn.neurotechedu.com/introtobci/.

Noroozi, F., Corneanu, C. A., Kamińska, D., Sapiński, T., Escalera, S., & Anbarjafari, G. (2018). Survey on emotional body gesture recognition.

Ren, & Zhi-peng. (2014). Body language in different cultures.

Shan, C. (2012). Learning human emotion from body gesture. In N. M. Seel (Ed.), Encyclopedia of the sciences of learning (pp. 1887–1889). Boston, MA: Springer US. doi: 10.1007/978-1-4419-1428-6_1905

Steinert, S., & Friedrich, O. (2020, Feb 01). Wired emotions: Ethical issues of affective brain–computer interfaces. Science and Engineering Ethics, 26(1), 351-367. doi: 10.1007/s11948-019-00087-2

Stephen Rainey, J. S., Hannah Maslen. (2020). When thinking is doing: Responsibility for bci-mediated action. AJOB Neuroscience, 11(1), 46-58. (PMID: 32009590) doi: 10.1080/21507740.2019.1704918

Wolkenstein, A., Jox, R., & Friedrich, O. (2018, 10). Brain–computer interfaces: Lessons to be learned from the ethics of algorithms. Cambridge Quarterly of Healthcare Ethics, 27, 635-646. doi: 10.1017/S0963180118000130