Turning negative into positives! Exploiting "negative" results in Brain-Machine Interface (BMI) research

Fabien Lotte[*,a] and Camille Jeunet[*,b] and Ricardo Chavarriaga[*,c] and Laurent Bougrain[*,d] and Dave E. Thompson[e] and Reinhold Scherer[f] and Md Rakibul Mowla[e] and Andrea Kübler[g] and Moritz Grosse-Wentrup[h] and Karen Dijkstra[i] and Natalie Dayan[j]

[*]Co-first authorship (these authors contributed equally)
[a] Inria, LaBRI (CNRS / Univ. Bordeaux / Bordeaux INP), France; [b] CLLE Lab (CNRS, Univ. Toulouse Jean Jaurès), France; [c] Chair in Brain-Machine Interface, École Polytechnique Fédérale de Lausanne, Geneva, Switzerland; [d] Univ. Lorraine, Inria Nancy Grand-Est / LORIA, Nancy, France; [e] Brain and Body Sensing Laboratory, Department of Electrical and Computer Engineering, Kansas State University, USA; [f] Brain-Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, United Kingdom; [g] Institute of Psychology, University of Würzburg, Germany; [h] Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna; [i] Radboud University, Donders Institute for Brain, Cognition and Behaviour; [j] Ulster University, Northern Ireland

ARTICLE HISTORY
Compiled November 21, 2019

ABSTRACT
Results that do not confirm expectations are generally referred to as "negative" results. While essential for scientific progress, they are too rarely reported in the literature - Brain-Machine Interface (BMI) research is no exception. This led us to organize a workshop on BMI negative results during the 2018 International BCI meeting. The outcomes of this workshop are reported herein. First, we demonstrate why (valid) negative results are useful, and even necessary for BMIs. These results can be used to confirm or disprove current BMI knowledge, or to refine current theories. Second, we provide concrete examples of such useful negative results, including the limits in BMI-control for complete locked-in users and predictors of motor imagery BMI performances. Finally, we suggest levers to promote the diffusion of (valid) BMI negative results, e.g., promoting hypothesis-driven research using valid statistical tools, organizing special issues dedicated to BMI negative results, or convincing institutions and editors that negative results are valuable.

KEYWORDS
Negative results, hypothesis, models, theory, publication, guidelines, BCI, BMI

1. Introduction

Negative results can be defined as "*results that do not confirm expectations*" [1]. For instance, they include results of an experiment or an evaluation of a new method

CONTACT Fabien Lotte. Email: fabien.lotte@inria.fr

that do not improve upon the state-of-the-art in terms of performance. They also include data from an experiment that did not confirm the hypothesis from which the experiment originated. While negative results are an unavoidable part of science, they tend to be less and less reported in scientific publications in general [1]. Indeed, Fanelli reported that "*The overall frequency of positive supports has grown by over 22% between 1990 and 2007*", meaning that there is an increasingly strong bias towards publishing positive results only.

Yet, it is widely acknowledged that valid[1] negative results are useful and even necessary for scientific progress [2–4]. Thus, negative results should certainly be useful for Brain-Machine Interface (BMI) research as well, especially since it is a young research field, in which most remains to be discovered and invented. Nonetheless, to the best of our knowledge, there have not been any dedicated publication discussing the status and value of negative results for BMI research. Moreover, according to the experience of the authors of this paper, BMI research also suffers, like other disciplines, from a bias towards reporting positive results only. In other words, currently, very few negative BMI research results seem to be published. This raises a number of questions: Why should we report negative results in BMI research? Are some unpublished negative results already relevant and useful? How to make sure that negative results are relevant? How to exploit and promote negative results in BMI research?

In order to answer these questions, we organized a workshop dedicated to negative results as part of the International Brain-Computer Interface Meeting 2018 in Asilomar, California, USA. The present paper aims at summarizing the findings and discussions from this workshop and the follow-up work that has been led by the participants. In particular, this paper aims at 1) presenting the value and usefulness of negative results, in science in general and in BMI research in particular (Section 2); 2) at presenting some existing negative results in BMI research - to illustrate concretely their importance and usefulness (Section 3); and 3) at proposing ways to promote the dissemination of negative BMI results, to enable the field to further progress (Section 4). Altogether, our goal with this manuscript is to contribute to the improvement of the scientific quality and diversity of BMI research, by encouraging the publication of relevant and valid scientific results, be these results positive or negative.

2. Negative results are valuable

As mentioned herein-above, negative results are results that can be useful and, most often, even necessary for scientific progress, e.g., to obtain new knowledge and develop new methods [2–4]. This is true for science in general, as depicted in Section 2.1, as well as for BMI in particular, as presented in Section 2.2 below.

2.1. Negative results are valuable in science in general

When a given experiment leads to negative results, whatever the scientific field, it may not be a priority to report them, particularly when the results do not seem to

[1]In this manuscript, when we mention scientific results, whether positive or negative, we generally (unless stated otherwise) assume that they are scientifically valid, i.e., that they originate from rigorous scientific studies, and are free of bias and confounding factors. Indeed, for both positive and negative results to be relevant and useful, they naturally first need to be scientifically valid. Thus, for the sake of conciseness, in the remainder of this manuscript, both the terms "negative results" and "positive results" refer to "valid negative results" and "valid positive results".

improve upon the state-of-the-art, e.g., in terms of performance. However, reporting them would spare other researchers unaware of them from wasting time, money and energy in conducting the same studies again. Actually, many labs at different points in time and space may work or may have worked on the same problems, all obtaining the same negative results unknowingly, if none of them reported those negative results. To ensure an efficient use of scientific resources and thus a more efficient scientific research worldwide, it is thus necessary to report negative results.

Additionally, any published scientific result, even sometimes a widespread one, might be a false positive (i.e., a study erroneously showed that a hypothesis was confirmed) or a false negative (a hypothesis was disproved when it was actually true) due to chance. In other words, as Ioannidis mentioned, some published results may be perpetuated fallacies, i.e., a piece of knowledge is repeatedly hold to be true by the community when it is actually not[2], while some genuine results may be left unconfirmed [5]. This is more likely to be case if this result was obtained on a small sample size population (see also next section for more details on that specific point). These errors cannot be corrected with a positive-only bias. For instance, replication studies obtaining different results than those already published are rarely reported, see, e.g., [6,7] for examples from the field of psychology. Moreover, and unfortunately, it should be acknowledged that scientific fraud or misconduct do happen, and even more often that we would like to believe, see, e.g., [8] for psychology. This also leads to published results that are actually false positives or negatives. All this thus makes the reporting of negative results necessary to identify such false positives or false negatives in science, and thus to contribute to make science self-correcting, as it should be [5,9]. Naturally, this also stresses the importance of good scientific practices.

In addition to enable us to correct science, by detecting published results that cannot be confirmed, negative results also enable us to refine science, and in particular to refine current models and theories from a given field. Indeed, negative results are necessary to identify where and when a method or a theory is valid, and where and when it is not. In other words, such results enable us to identify the scope of application, the contexts in which some methods will be useful, and the boundaries of some models and theories.

Finally, it should be stressed that failure is part of the research process, where substantial progress is achieved by trials-and-errors, and where scientists explore uncharted areas of knowledge. Moreover, science targets discovery and the generation of new knowledge, which requires to take risks. Such process inevitably leads to some failure to improve beyond the state-of-the-art, and to unexpected results - thus including numerous negative results. As scientists, we should learn from these failures and thus from negative results, to push our fields forward [10].

2.2. Negative results are valuable in BMI in particular

As for any scientific research field, negative results can thus be valuable for BMI research, for the reasons mentioned above. Additionally, some properties of BMI research make negative results even more valuable and necessary for this field.

First, there is a large between-user and within-user variability observed in BMI, in terms of brain activity patterns, BMI performances, signal processing method ef-

[2]Well known historical examples of perpetuated fallacies include when physicists believed the Earth to be flat or that the Sun orbited around the Earth.

ficiency or in terms of user learning, among other [11–14]. It means that a method that has proven efficient in a given context might not be suitable in another one. It also means that some results obtained on a given set of users or with a given signal processing or feedback method, for instance, may be valid only for this population and method, and might not be adequate for other users or with other methods. For instance, in [12], the authors have revealed that the Filter Bank Common Spatial Patterns (FBCSP) method –a widely used and efficient signal processing method for EEG-based BMIs– that has won several BMI competitions [15], can indeed perform much better than CSP on many data sets, but also significantly worse on some other data sets, at least when using cross-validation for performance evaluation. Similarly, while many classification methods seem to be very efficient for offline BMI analysis, the methods actually used for online BMI are generally different, and often simpler [13]. Another example of the importance of discussing alternative interpretations is the case of the 'rotational dynamics' observed in intracranial recordings in pre-motor areas and their link to behaviour [16,17]. These examples highlight the fact that we need negative results in order to identify in which context a given method is the most suitable.

Second, BMI studies are most often associated to small sample sizes, generally between 10 to 20 participants in EEG and as few as 2 in intracranial BMIs. These small sample sizes are likely to result in underpowered studies, which can, on the one hand, lead to false negative/positive outcomes. Moreover, on the other hand, the "statistically significant" results from underpowered studies have substantially lower positive predictive value, i.e., likelihood of reflecting an actual difference [18]. Furthermore, small sample sizes tend to exaggerate effect sizes through the "winner's curse" [19]. This exaggeration is dangerous as it can lead to an overstatement of clinical importance as well as leading future investigators to design further underpowered studies [18]. We thus need negative results being reported to confirm or disprove published results from underpowered BMI studies and provide evidence of the adequacy –or lack thereof– of previously accepted approaches.

Last but not least, BMI research is still a relatively young field, and as such a field still lacking models and theories to explain the results observed or to guide the choice of particular methods. Indeed, while we (the BMI research community) do have experimental results about some factors explaining, e.g., performance variabilities between or within users or about how users learn to control BMIs [14,20,21], we have very few actual theories or models explaining why it is so, and how to influence that in practice and precisely[3]. We also lack models or theories explaining why a given machine learning or signal processing algorithm works well on some users or some BMI paradigms, but not so well on some others. Yet, models and theories are essential to structure, conceptualize and guide a research endeavour towards further progress [26]. We thus need both positive and negative results to build, refine and contextualize such models and theories. We indeed cannot build such theories with positive results only, as we need to compare various theories, identify which ones are valid and which ones are not, and identify their scope as well as the limits of their application. This, again, requires negative results (as well as positive ones of course).

In summary, reporting negative results could enable everyone to save time and resources, and could help us to identify false negatives and false positives in published results, to validate and refine existing knowledge and tools. This is particularly use-

[3]Some works in that direction include [22,23] for BMI or [24,25] for Neurofeedback

ful for BMI research in which the large variabilities observed and the typically small sample sizes used makes already published results in need for being confirmed or disproved, possibly with negative results. Finally, BMI research being critically lacking models and theories, it needs both positive and negative results to build and refine them.

3. Examples of useful negative results in BMI

As mentioned in the introduction, negative results can be defined as "*results that do not confirm expectations*". In this section, we introduce four concrete examples of negative results in the field of BMIs in order to illustrate that they can be useful to the BMI community. These examples highlight results that are not significant, results that contradict the literature, incomplete results and unexpected results.

Some obtained results may seem to contradict previous publications. Except for replication studies, many parameters can change between two studies, starting with the influence that the experimenter can have on the users understanding, motivation and confidence [27,28]. Indeed, especially in studies involving human beings, users and experimenters can have a notable influence on the results. As an example, Rimbert et al. [29] have conducted an experiment whose goal was to confirm the effectiveness of the use of subjective questionnaires, such as the Motor Imagery Questionnaire Revised-Second Edition (MIQ-RS), to estimate the performance of a Motor Imagery (MI)-based BMI. Predicting a subject's ability to use a BMI is one of the major issues in the BMI domain. To be relevant, BMI applications should be able to adapt to the needs and expectations of the user. The authors recorded EEG signals from 35 healthy volunteers during MI-BMI use. The subjects had previously completed the MIQ-RS questionnaire. They conducted an offline analysis to assess the correlation between the questionnaire scores related to Kinesthetic and Motor imagery tasks and the BMI performances, using four different classification methods. The results revealed that BMI performance correlated with participants' habits and frequency of manual activities practice. Nonetheless, no significant correlation was revealed between BMI performance and the MIQ-RS scores, unlike previously reported studies [30,31]. However, as mentioned above, there were a few differences between these three studies. Thus, these results should not necessarily be considered as negative results just because they contradict previous studies. Gathering results from several studies targeting the same factors will help to understand the factors actually having an influence. Ideally, publishing the protocol details (hardware and software used, environmental conditions, number and type of experimenters, ...) and sharing the code used and the data collected would allow other scientists to properly replicate studies, to enlarge the sample size, and to identify possible confounding factors.

Nevertheless, replication studies can also be difficult to publish. In one recent study, the authors from Kansas State University attempted to replicate previous demonstrations of BMIs for affective classification using the International Affective Picture System (IAPS) as stimuli [32]. Participants rated each picture for different affective dimensions (valence, arousal, and dominance) using a self-assessment manikin (SAM) [33]. Classification was performed using literature-based features and techniques. The initial results seemed encouraging, with some users achieving greater than 80% accuracy on multiple axes in a binary high/low classification approach; most participants had performance statistically significantly above 50%. However, upon examining the data, the investigators discovered substantial response bias for each participant. The

response bias was so severe that an unskilled classifier, which simply guessed the most common class without considering EEG data, performed as well as the proposed system. Despite the important observation that consideration of response bias is required, and that many prior studies have not taken it properly into account, the team has had difficulties publishing the results because they are not statistically significant by the team's own metric.

Some other studies can clearly show an effect produced by a given method, however without being fully able to explain it. In these cases, researchers can be reluctant to publish their results. The BMI domain is a multidisciplinary field requiring expertise in biology, physiology, signal processing, machine learning, computer science and psychology, among others. Therefore, suitable interpretations could also come from the community, after such uninterpreted results have been published. Publishing these unexplained results could thus also help to increase our knowledge. For example, Rimbert and al. [34] evaluated the influence of an hypnotic condition on Event Related Desynchronization/Sychronization (ERD/ERS) patterns during a kinesthetic motor imagery (KMI) task. Indeed, hypnotic inductions using Ericksonian suggestions can make the user feel simultaneously more relaxed and more focused on the mental task, and therefore could be used to increase BMI performance. To investigate this issue, 19 right-handed healthy subjects performed a KMI of the right hand during two randomized sessions: in normal and hypnotic conditions. Their results suggested that the state of hypnosis shortened the ERD phase in the sensorimotor frequency bands, assuming a change in the activation of the motor cortex during the hypnotized state and thus a worst detection of the kinesthetic motor imagery. These results prompted the authors not to recommend using hypnosis for motor imagery-based BMI applications, even though the underlying psychological and physiological phenomenons are not fully understood yet. Further studies on the effects of alteration of consciousness methods such as meditation and sophrology may help identify the origin of this finding.

In the field of BMIs for individuals with amyotrophic lateral sclerosis (ALS), the reporting of negative results has been essential for recent progress. After initially promising results in patients who had not yet entered the completely locked-in (CLIS) state of ALS [35], the failure to establish BMI-based communication in CLIS-ALS patients prompted Kübler et al. to formulate their influential hypothesis on the extinction of goal-directed thinking in complete paralysis [36]. Only recently, various research groups have started to carefully document which skills are (not) preserved once patients enter the CLIS. Specifically, Okahara et al. reported evidence for a CLIS-ALS patient to retain the capacity for command-following. They could not, however, establish goal-directed communication [37]. Goal-directed communication with an ALS patient was recently reported by Han et al. [38], but this result could not be reproduced in a follow-up study four months later. As there is no reason to assume that ALS ceases to progress once patients have entered the CLIS, and recent evidence pointing to a collapse of the frequency of the α-rhythm in long-term CLIS-ALS patients [39], well-documented negative results are essential to foster a deeper understanding regarding which skills CLIS-ALS patients retain –and for how long.

Interestingly enough, whether CLIS-ALS patients can in fact use a BMI for communication was actually a recent topic of debate in the BMI community. Indeed, two different groups obtained different results on the same fNIRS (functional Near Infrared Spectroscopy) data of CLIS-ALS patients. One group obtained a positive result (CLIS-ALS patients can use a BMI) while the other obtained a negative result (CLIS-ALS patients cannot use a BMI) when re-analyzing the same data [40–42][43]. This thus again illustrates the need to report both positive and negative results, to clarify what

such patients can actually do or not.

4. Encouraging and promoting the dissemination of negative results

In order to overcome the lack of negative results in the BMI literature, it is important to consider why negative results go under-reported. If academics are simply unfamiliar with the extent and consequences of this issue, then raising awareness of the problem may be sufficient. However, as Ioannides et al. point out, even well-intentioned scientists may be subject to biases that lead to the prioritizing of positive over negative results, in particular when these biases align with existing pressures in academia [44]. For instance, given the pressure to publish, authors perceiving lower probabilities of being able to publish negative results, may focus their efforts elsewhere than writing up a study with a negative result. The pressure to publish in high impact factor journals, which usually do not, or very rarely, publish negative results [45] (even though the same journals may claim such results are necessary [2]) could also implicitly train scientists to ignore negative results and ignore the value of failure in science. Negative results may also appear as much more difficult to interpret. Indeed, there could be many reasons for the observed negative results whereas positive results may simply help to confirm an hypothesis. Finally, scientific competition might sometimes lead some researchers not to report negative results, so that others would also waste their time on the same problem, thus making them less likely to publish first an important positive result. While it is unclear whether this problem affects the BMI community, and if it does, to which extent, it might still contribute to make negative results under-reported.

In addition to educating researchers and raising awareness of the issue, we can thus consider opportunities to set up incentives to encourage scientists to report their results, whether positive or negative. In the sections below we break down proposed solutions into suggestions at the methodological level, the publication level, and the community/institutional level.

4.1. At a methodological level

Scientists take pride in the self-correcting nature of science, however notions that remain unchallenged become dogma. The scientific process allowed disproval of long-standing beliefs like the geocentric universe theory or the idea of spontaneous generation. However, this process can only be achieved if we accept the worthiness of results that do not conform with the prevalent underlying hypothesis. Even though non-conformists are no longer burned to death, studies that fail the expected result are often reduced to ashes by the difficulty of making them part of the scientific debate.

Systemic factors contribute to the disregard of the so-called negative results. As mentioned above there is a bias in the publication system towards success stories that support the feeling of progress in the field. Hence, scientific journals and media relish on demonstrations of BMI systems irrespective of how solid the evidence of their claims is. In addition, there is a general misconception that there is nothing to be learned from these so-called negative results. This is even more pronounced in fields like BMI that rely heavily on empirical evaluations due to a lack of well-proven models of the brain processes these systems are trying to decode.

We argue that talking about positive or negative results is a distracting fallacy. So much of scientific methodology entails formalized training, however many BMI

researchers are able to make it through their entire careers without specific training in hypothesis development and statistical analysis, which explains in part why so many studies attempt to draw conclusions from under-powered studies. If you are seeking a pattern, there can be a tendency to dishonour the pattern that is truly there. Nonetheless, certain techniques may mitigate malpractice including: familiarization with critical thinking approaches to the scientific method [46], challenging assumptions through feedback from peers and applying non-biased formulations for predicting different experimental outcomes. The more mutually exclusive hypotheses are, arguably the greater the fruitfulness of the scientific enquiry. With widespread adoption of valid hypothesis testing, the stigma towards negative results may diminish. The granularity and testability of a study hypothesis directly influences the explanatory power and validity of the results, irrespective of whether the results are negative or positive. Overall, the success of an experiment should be measured by its ability to distinguish between different contingencies rather than attaining the anticipated outcomes without a supporting valid hypothesis.

As a community, we should foment studies that lead to better understanding and can potentially improve the state-of-the-art, irrespectively of whether the outcomes fulfil the initial hypothesis. This requires a more formal approach for designing experiments [47,48]. Data analysis methods should be appropriately chosen to assess the strength of the obtained results. This is particularly important in the BMI field where most studies involve small populations in one or just a few sessions. These studies are often statistically underpowered and therefore even positive results obtained in these conditions should not be considered sufficient evidence to draw solid conclusions. Importantly, this need is not exclusive to BMI but also pertains to closely related fields like neuroscience and cognitive psychology [49,50]. Nonetheless, specific characteristics and constraints of BMI scenarios make this approach more difficult. In particular, gathering data is highly demanding in terms of time and resources and the large variability within and across subjects makes it difficult to obtain meaningful results from such small sample sizes.

Nonetheless, several aspects can be taken into account to improve experimental methods design in BMI in order to obtain meaningful results. These include, among others, the choice of the experimental conditions to be tested, the types of subjects that are included in the study and the metrics used to assess performance. For sake of space, we briefly discuss some general recommendations (relevant material is found in previous publications on this topic [47,48,51–53]).

As mentioned above, one of the biggest limiting factors is the small sample size that take part in BMI studies. Given this constraint, an alternative is to privilege long-term studies over these small populations [54]. Such studies, although demanding, give a better assessment of how a BMI will work in its intended operational conditions [55,56]. Nonetheless, care should be taken to choose appropriate experimental protocols and analysis approaches for studies with small N, and results should not be interpreted as generalizable to larger populations. Additionally, studies should include relevant control conditions and tests to account for potential confounds that may exist. These include proper assessment of chance level and choice of statistical sets suitable for the chosen sample size [57–61].

In conclusion, consistent obtention of meaningful results, either confirming or refuting the initial hypothesis, would be possible by improving the experimental design. Hence, efforts should be made to incentivize this. One of them is the pre-registration of studies before they are performed. This will encourage devoting more efforts to design experimental protocols and data analysis methods (see below). Also, mentors should

take special care on teaching young researchers the inherent value of learning from failed, well-designed experiments. Last but not least, the gatekeepers of scientific communication –journal editors, conference chairs and hiring committees– should clearly adopt the idea that works should be evaluated based on the solidity of their results and the extent they are able to support valid conclusions. These last points will be further developed in the following sections.

4.2. At a publication level

Encouraging the reporting of negative results can also be made at a publication level, notably by encouraging or easing the submission/publication of manuscripts reporting valid negative results.

For instance, one way of doing so would be to organize special issues in scientific journals, and/or special sessions (with talks and/or posters) in scientific conferences, that are dedicated to report negative BMI results. This way, scientists who obtained negative results in a valid and rigorous way would not be afraid or reluctant to submit them as there would be a dedicated venue to do so. Such an approach would also provide more visibility to negative results, hopefully showing them under a more positive light, thus also favoring their future reporting. Hopefully, after a number of special issues and special sessions to bootstrap the process, and a number of negative results being published, publishing new negative BMI results should become natural and accepted, and thus should become a standard practice, as it should be.

At the publication level, reporting negative results is only partially in the hands of the researcher. Editors of journals and reviewers are often also biased toward positive results, and, thus, reluctant to accept reports on negative results. With the popularization of preprints services (e.g., arXiv.org and bioRxviv.org), there is a viable option for authors to make their research publicly available. Importantly, making a manuscript available as a preprint, in most cases, does not prevent subsequent publishing in a journal, as a large number of publishers and journals allow preprints [4] [5]. These preprints are indexed by search engines and allow others to become aware of the work even if it has not been peer reviewed.

The study to be reported, however, is entirely in the hand of the researcher. If we follow systematically a line of thought with a sequence of thoroughly designed experiments, reporting of negative results should not be a problem as the message lies in the line of thought. The line of logic has to be made transparent and as simple as possible. Specifically in the so-called hard sciences, such as molecular biology, this accumulative method of inductive inference is systematically used and taught, and was termed strong inference (p. 347, [62]). The approach of inductive inference goes back to Francis Bacon [63] (cf. [53,62]). According to Platt, four steps are crucial for strong inference. Taub describes three additional steps that were formulated by T.C. Chamberlin [53,64]; Table 1 summarizes those steps and lists an example from the BMI field.

It is worth thinking of as many alternative hypotheses as possible which reflects the whole logical structure of a problem. The deducted sequence of experiments allows for a successive approximation (p. 111, [53]) to the phenomenon and guides interpretation of results. Those steps are familiar to all of us, but the difference arises from their

[4] `https://en.wikipedia.org/wiki/List_of_academic_journals_by_preprint_policy`. Note that for some journals it matters which copyright license you publish your preprint under (e.g., at the time of writing, IOP publishing does not allow preprints that have been published under a Creative Commons license.)

[5] `http://www.sherpa.ac.uk/romeo/search.php`.

Table 1. The seven/four steps constituting strong inference according to [53] and [62]

Step	Content	Example for BMI
1	Select a phenomenon	BMI inefficiency - a modulation of brain activity (e.g., SMR) cannot be detected by the BMI
2	Ask a question about the phenomenon	Is BMI inefficiency influenced by age?
3	Devise alternative hypothesis (Platts step 1)	Is BMI inefficiency influenced by neurodegeneration? By intelligence?
4	Define and design a crucial experiment (Platts step 2)	Several BMI sessions with different age groups (independent variable); measure performance (dependent variable); conduct a power analysis to define the sample size
5	Carry out the experiment (Platts step 3)	Ensure adherence to the experimental procedure
6	Recycle the procedure, include sub-hypotheses (Platts step 4)	Sub-hypothesis: is performance moderated by the ability to focus attention
7	Select for consideration a new phenomenon that emerged during the process	Is SMR-BMI performance influenced by the default EEG spectrum?

systematic application. Applying this method more often and stringently in the field of BMI would increase the quality of studies and facilitate their subsequent publication independently of the results being positive or negative.

Finally, scientific advancement is only plausible when findings are deemed to be of sufficient credibility. In the face of the systemic publish or perish culture, the likelihood of scientific misconduct or overlook in research design, administration and interpretation by research groups is heightened. This can lead to mass publication of untrustworthy results, which is arguably demonstrated by the replication crisis [65]. The evolving open science research culture is bringing the robustness of scientific methodologies to the forefront, in order to reinstate trust in scientific findings. Registered reports are an example of such efforts, whereby, research articles with the proposed protocol and analyses are written up and peer reviewed by journals prior to the research being conducted. Acceptance of the protocol at this stage by journals guarantees publication irrespective of whether the results are positive or negative. Currently 187 journals (`https://cos.io/rr/`) use registered reports publishing format as a regular submission option or as part of a special issue, however no BMI dedicated journals have implemented this submission format (although BMI research is published far and wide due to the multidisciplinary nature of the field). In addition, pre-registration of research designs is an additional effort, whereby researchers can publicly disseminate a data analysis plan prior to observing research results, to enhance scientific rigour and mark a priori explanatory versus exploratory hypotheses investigations. Another interesting initiative is the Open Science Framework (OSF) [66] (see also

http://osf.io), which provides, for free and open-source, online resources and tools to share research, i.e., to share protocols, codes, documents and data (among other), and to work collaboratively on them. Such initiatives also provide interesting guidelines about how to conduct studies, report them and share the data, see, e.g., OSF guidelines for M/EEG in [67] or the dedicated IEEE standards project for In Vivo Neural Interfaces (https://standards.ieee.org/project/2794.html). Note however, that these guidelines are typically designed for open-loop experiments whereas BCI research is fundamentally close-loop. Altogether, these initiatives naturally promote replicability and good scientific practices, both for positive or negative results. Successful adoption of the open science culture can thus enhance scientific credibility of publications and hence eliminate the systemic bias against negative results, and mitigate bad research practices like selective reporting of results or low statistical power.

4.3. At a community and institutional level

Even if changing researchers' mindset about negative results may take some time, we could, as a scientific community, lead actions to promote more reasonable practices. In the following paragraphs, we suggest four levers of action, that could of course be completed by initiatives from the members of the BMI community. These levers are the following: promoting collaboration between teams, creating a shared platform, producing papers identifying open research questions and convincing institutions and editors about the necessity to publish negative results.

First, in order to advance the knowledge in our field, we should (as extensively mentioned herein-above) perform replication studies, to confirm or disprove previously published results (both positive and negative). To do so, we could for instance create consortia between research teams, with BMI researchers who would collaborate by sharing their protocols and BMI implementations, and replicating experiments led by other teams of the consortium they belong to. Such a practice would be beneficial in several ways. It would on the one hand enable the replications to be led in different contexts, thus getting rid of potential biases, and on the other hand enable us to increase the statistical power of our experiments by increasing the number of participants. It could also enable research teams to follow-up on previous experiments led by other teams.

In the same vein, in the age of internet, implementing a collaborative online service that facilitates discussion between researchers would be a reasonable approach. This online service could, among others, include a forum as well as a platform on which researchers could share paradigms and methods, share data and code, and store relevant papers. The Open Science Framework mentioned above is one tool that can make that possible, and thus, that should probably be considered by the BMI research community [66]. Indeed, scientific discussions and exchanges of ideas are engines that drive science forward. The BMI field is, by definition, interdisciplinary and the diversity of the knowledge areas involved is necessary to be successful. Concise and clear communication is essential. The advantage of such an approach is that discussions, ideas and results –including negative results– are preserved for the future. Querying the accumulated data would support researchers and allow them to progress faster. However, such an approach can only be successful if the community agrees to invest time and check plausibility and integrity of the contained information. So there is no short term gain but a long term investment.

Furthermore, as mentioned in Section 4.1, having valid research hypotheses and clear research questions would help to design experiments the outcome of which should be useful, be the results positive or negative. The BMI research community could favor this approach by publishing, possibly regularly (e.g., every couple of years) papers summarizing open questions and challenges in BMI research or on sub-parts of BMI research (e.g., open questions on EEG signal processing or on clinical BMI applications). Such papers could also provide possible alternative hypotheses targeting the current state of knowledge in BMIs, and that should be tested by the community. Such an approach would provide clear and valid research questions and hypotheses to be tested by BMI scientists, which could in turn favor relevant and insightful outcomes, be they positive or negative. Moreover, it would also encourage hypothesis-driven research, that aims at answering research questions deemed outstanding by the community, instead of being driven by the current bias towards seeking positive results only.

Nonetheless, the publication of such papers cannot be done without the consent and engagement of scientific editors. There, it seems that a change of paradigm is needed. Currently, the results novelty and the research originality seem most important for getting a paper accepted. Novelty, in an interdisciplinary field, however, is very challenging. What is novel for a biologist may not be new for a computer scientist and vice versa. Some journals started putting focus on the soundness of the methods rather than on the reviewers' rating of the novelty of the approach. Such an approach allows publication of positive and negative results. If we want this sound approach to become mainstream, we should –as far as possible– get involved in editorial boards and push towards this direction. Then, if we want researchers to actually perform replication studies and publish negative results, we need first to ensure that it will not be detrimental for their career. More precisely, replication studies should be considered necessary to consolidate a theory, and not a loss of time; while the publication of negative results should be considered an advance in the field rather than a failure. For researchers to adopt this mindset, they should explicitly be encouraged, both by funding institutions (e.g., the European Research Council and national research agencies) and by academic institutions (e.g., University councils). We, once again as a scientific community and hopefully with the support of the BCI Society, may advise these institutions to recognize negative results and replication studies as being just as valuable as any other result/study. Finally, through our involvement in hiring committees, we could also promote applicants who have contributed to the field by leading replication studies and publishing negative results.

5. Conclusion

In this paper, we have described how (valid) negative results, which represent results that do not confirm our expectations, are too rarely reported in scientific publications, including in BMI publications. Nonetheless, such negative results can and are often very valuable and even necessary for scientific progress. We mentioned how they can indeed enable other labs to save time, allow to confirm or disprove current pieces of knowledge that may be false positives/negatives –which is particularly likely in BMI research given the small sample sizes used and the data variability–, or help us building and refining accurate and comprehensive models and theories, models and theories that the BMI field is still lacking.

In order to further illustrate the relevance of negative results, we presented a few actual examples of negative BMI results, which are useful to deepen our understand-

ing of BMI technology and usage. These examples included first a demonstration that questionnaires measuring MI abilities cannot always predict MI-BMI performances, second an illustration about the fact that hypnosis –which is hypothesized to increase attentional abilities– actually is detrimental to MI-BMI performances. The third example suggests that decoding emotions from EEG when using the IAPS for emotion elicitation may not achieve better-than-chance performances when considering class imbalance, while the fourth example illustrates the fact that negative results are necessary to understand what CLIS-ALS patients can or cannot do, e.g., regarding the use of a BMI.

In order to promote the diffusion of negative results, we suggested a number of levers, at different levels. At the methodological level, we suggested to improve our BMI experimental designs, by encouraging hypothesis-driven research, with clear research questions, and with suitable statistical tools and power, to ensure that any result from such studies would be valuable, whether it is positive or negative. We also suggested to promote the value of learning from "failed" experiments. At the publication level, we suggested to organize conference special sessions and journal special issues dedicated to negative results, to bootstrap the process of publishing them. We also recommended the use of pre-print publishing of negative results; to publish and present work based on principles such as strong inference and to pre-register experimental protocols. Finally, at the community and institutional levels, we argued for increasing collaborations between teams, with a shared research platform (e.g., with OSF), where all kinds of results could be shared. In general, following open science principles should increase the credibility, validity and relevance of published results, both positive and negative ones. We also recommended to produce opinion papers identifying current open BMI research questions, as well as to convince institutions and scientific editors of the value of negative results, so that these can be published and valued in a scientific career.

In the end, what we tried to convey was that for BMI research to progress, it should not matter so much whether the results obtained and reported are positive or negative. Rather, it should matter whether such results are valid and relevant, i.e., whether they notably originate from experiments and studies that followed rigorous and sound scientific practices and methodologies, are sufficiently powered, answer a clear hypothesis or research question, and are supported by appropriate statistical analyses. Indeed, as mentioned earlier, the use of terms such as "negative" and "positive" is probably misleading, as there is nothing negative about a so-called negative result (when it is valid of course). It is time to forget that vocabulary, and aim to obtain and publish "valid results". Thus, with this paper, we aim at raising awareness in the BMI community on the need to report all their valid results, whether they are positive or negative. This paper is only a first step towards this goal, and we intend and encourage the initiation of a few others, in line with the suggestions mentioned above. To do so, we will need the help of the whole BMI community, to organize related events and to encourage the publication of negative results. For instance, it would be very useful to produce, in collaboration, publication, experimental and reviewing guidelines for BMI research. This approach has recently been initiated in the neurofeedback field [68,69]. This would favor relevant and valid BMI experimental designs, and thus useful and valid results, be they positive or negative. We also plan to launch a special issue dedicated to reporting negative BMI results to further promote them. We hope that the BMI community will join us in this endeavour.

Acknowledgement(s)

The authors would like to acknowledge all the participants of the workshop at the Asilomar BCI meeting 2018, which is at the origin of the present paper.

Funding

F.L. was partly supported by the European Research Council, with project Brain-Conquest (grant ERC-2016-STG-714567) and the French National Research Agency with project REBEL (grant ANR-15-CE23-0013-01). R.C. was supported partially by the IEEE Brain Initiative and the IEEE Standards Association Industry Connections Program[6]. C.J. was partially supported by the EPFL-Inria International Lab as well as by the Swiss National Foundation (SNF) for scientific research.

References

[1] Fanelli D. Negative results are disappearing from most disciplines and countries. Scientometrics. 2012;90(3):891–904.
[2] Knight J. Negative results: Null and void. Nature. 2003;422:554–555.
[3] Vance E. Nurture negatives. Nature. 2017;552(7685):302–302.
[4] Granqvist E. Looking at research from a new angle: why science needs to publish negative results. Elsevier Publishing Ethics, Accessed October. 2015;4:2016.
[5] Ioannidis JP. Why science is not necessarily self-correcting. Perspectives on Psychological Science. 2012;7(6):645–654.
[6] Yong E. Replication studies: Bad copy. Nature News. 2012;485(7398):298.
[7] Smith NC. Replication studies: A neglected aspect of psychological research. American Psychologist. 1970;25(10):970.
[8] John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological science. 2012;23(5):524–532.
[9] Alberts B, Cicerone RJ, Fienberg SE, et al. Self-correction in science at work. Science. 2015;348(6242):1420–1422.
[10] Firestein S. Failure: Why science is so successful. Oxford University Press; 2015.
[11] Krusienski DJ, Grosse-Wentrup M, Galán F, et al. Critical issues in state-of-the-art brain–computer interface signal processing. Journal of neural engineering. 2011;8(2):025002.
[12] Jayaram V, Barachant A. Moabb: trustworthy algorithm benchmarking for bcis. Journal of neural engineering. 2018;15(6):066011.
[13] Lotte F, Bougrain L, Cichocki A, et al. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. Journal of neural engineering. 2018; 15(3):031005.
[14] Jeunet C, N?Kaoua B, Lotte F. Advances in user-training for mental-imagery-based bci control: Psychological and cognitive factors and their neural correlates. Progress in brain research. 2016;228:3–35.
[15] Ang KK, Chin ZY, Wang C, et al. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. Frontiers in neuroscience. 2012;6:39.
[16] Lebedev MA, Ossadtchi A, Mill NA, et al. What, if anything, is the true neurophysiological significance of "rotational dynamics"? bioRxiv. 2019 jan;:597419Available from: `http://biorxiv.org/content/early/2019/04/12/597419.abstract`.

[6] `https://standards.ieee.org/industry-connections/neurotechnologies-for-brain-machine-interfacing.html`.

[17] Churchland MM, Cunningham JP, Kaufman MT, et al. Neural population dynamics during reaching. Nature. 2012;487(7405):51–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22722855{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3393826.

[18] Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013 May;14(5):365–376. Available from: https://www.nature.com/articles/nrn3475.

[19] Zllner S, Pritchard JK. Overcoming the Winners Curse: Estimating Penetrance Parameters from Case-Control Data. The American Journal of Human Genetics. 2007 Apr; 80(4):605–615. Available from: http://www.sciencedirect.com/science/article/pii/S0002929707610970.

[20] Grosse-Wentrup M. What are the causes of performance variation in brain-computer interfacing? International Journal of Bioelectromagnetism. 2011;13(3):115–116.

[21] Lotte F, Larrue F, Mühl C. Flaws in current human training protocols for spontaneous brain-computer interfaces: lessons learned from instructional design. Frontiers in human neuroscience. 2013;7:568.

[22] Kübler A, Blankertz B, Müller K, et al. A model of BCI-control. In: Proceedings of the 5th International Graz BCI conference; 2011.

[23] Jeunet C, N'Kaoua B, Lotte F. Towards a cognitive model of MI-BCI user training. In: Proceedings of the International Graz BCI conference; 2017.

[24] Gruzelier J. A theory of alpha/theta neurofeedback, creative performance enhancement, long distance functional connectivity and psychological integration. Cognitive processing. 2009;10(1):101–109.

[25] Wood G, Kober SE, Witte M, et al. On the need to better specify the concept of control in brain-computer-interfaces/neurofeedback research. Frontiers in systems neuroscience. 2014;8:171.

[26] Kuhn TS. The structure of scientific revolutions. University of Chicago press; 2012.

[27] Wood G, Kober SE. Eeg neurofeedback is under strong control of psychosocial factors. Applied psychophysiology and biofeedback. 2018;43(4):293–300.

[28] Roc A, Pillette L, N'Kaoua B, et al. Would motor-imagery based bci user training benefit from more women experimenters? In: Proceedings of the International Graz BCI conference; 2019.

[29] Rimbert S, Gayraud N, Bougrain L, et al. Can a subjective questionnaire be used as brain-computer interface performance predictor? Frontiers in Human Neuroscience. 2019;12:529. Available from: https://www.frontiersin.org/article/10.3389/fnhum.2018.00529.

[30] Vuckovic A, Osuagwu BA. Using a motor imagery questionnaire to estimate the performance of a braincomputer interface based on object oriented motor imagery. Clinical Neurophysiology. 2013;124(8):1586 – 1595. Available from: http://www.sciencedirect.com/science/article/pii/S1388245713001120.

[31] Marchesotti S, Bassolino M, Serino A, et al. Quantifying the role of motor imagery in brain-machine interfaces. Scientific Reports. 2016 04;6:24076.

[32] Lang PJ, Bradley MM, Cuthbert BN. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. The Center for Research in Psychophysiology, University of Florida, FL, USA; 2008. Technical report a-8.

[33] Bradley MM, Lang PJ. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry. 1994;25(1):49–59.

[34] Rimbert S, Zaepffel M, Riff P, et al. Hypnotic state modulates sensorimotor beta rhythms during real movement and motor imagery. Frontiers in Psychology. 2019;10:2341.

[35] Birbaumer N, Ghanayim N, Hinterberger T, et al. A spelling device for the paralysed. Nature. 1999;398(6725):297–298.

[36] Kübler A, Birbaumer N. Brain-computer interfaces and communication in paralysis: Extinction of goal directed thinking in completely paralysed patients? Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology. 2008; 119(11):2658–2666.

[37] Okahara Y, Takano K, Nagao M, et al. Long-term use of a neural prosthesis in progressive paralysis. Scientific Reports. 2018;8(1):16787.
[38] Han CH, Kim YW, Hyun SS, et al. Electroencephalography-based endogenous brain–computer interface for online communication with a completely locked-in patient. Journal of Neuroengineering and Rehabilitation. 2019;16(1):18.
[39] Hohmann M, Fomina T, Jayaram V, et al. Case series: Slowing alpha rhythm in late-stage als patients. Clinical Neurophysiology. 2018;129(2):406–408.
[40] Spüler M. Questioning the evidence for BCI-based communication in the complete locked-in state. PLoS biology. 2019;17(4):e2004750.
[41] Chaudhary U, Pathak S, Birbaumer N. Response to:questioning the evidence for BCI-based communication in the complete locked-in state. PLoS biology. 2019;17(4):e3000063.
[42] Scherer R. Thought-based interaction: Same data, same methods, different results? PLoS biology. 2019;17(4):e3000190.
[43] Abbott A. Prominent german neuroscientist committed misconduct in brain-reading research. Nature News. 2019;Available from: https://www.nature.com/articles/d41586-019-02862-4.
[44] Ioannidis JPA, Munaf MR, Fusar-Poli P, et al. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends in Cognitive Sciences. 2014 May;18(5):235–241. Available from: http://www.sciencedirect.com/science/article/pii/S1364661314000540.
[45] Matosin N, Frank E, Engel M, et al. Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. Disease Models & Mechanisms. 2014;7(2):171.
[46] Blachowicz J. How Science Textbooks Treat Scientific Method: A Philosopher's Perspective. The British Journal for the Philosophy of Science. 2009 jun;60(2):303–344.
[47] Jeunet C, Debener S, Lotte F, et al. Mind the traps! Design guidelines for rigorous BCI experiments. In: Brain-computer interfaces handbook: Technological and theoretical advance. Taylor & Francis; 2018.
[48] Chavarriaga R, Fried-Oken M, Kleih S, et al. Heading for new shores! overcoming pitfalls in BCI design. Brain-Computer Interfaces. 2017;4(1-2):60–73.
[49] Alger BE. Hypothesis-testing improves the predicted reliability of neuroscience research. bioRxiv. 2019;.
[50] Muthukrishna M, Henrich J. A problem in theory. Nature Human Behaviour. 2019 mar;3(3):221–229.
[51] Varoquaux G, Raamana PR, Engemann DA, et al. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage. 2017;145(August 2015):166–179.
[52] Brouwer AM, Zander TO, Van Erp JB, et al. Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. Frontiers in Neuroscience. 2015;9:136.
[53] Taub E. What Psychology as a Science Owes Neal Miller: The Example of His Biofeedback Research. Biofeedback. 2010 sep;38(3):108–117. Available from: http://www.aapb-biofeedback.com/doi/abs/10.5298/1081-5937-38.3.108.
[54] Smith PL, Little DR. Small is beautiful: In defense of the small-N design. Psychonomic Bulletin and Review. 2018;25(6):2083–2101.
[55] Saeedi S, Chavarriaga R, Millán JdR. Long-Term Stable Control of Motor-Imagery BCI by a Locked-In User Through Adaptive Assistance. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2017 apr;25(4):380–391. Available from: http://ieeexplore.ieee.org/document/7801122/.
[56] Sellers EW, Ryan DB, Hauser CK. Noninvasive brain-computer interface enables communication after brainstem stroke. Sci Transl Med. 2014 oct;6(257):257re7. Available from: http://dx.doi.org/10.1126/scitranslmed.3007801.
[57] Kass RE, Caffo BS, Davidian M, et al. Ten Simple Rules for Effective Statistical Practice. PLoS Comput Biol. 2016 jun;12(6):e1004961. Available from: http://dx.doi.org/10.1371/journal.pcbi.1004961.

[58] Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science. 2014 nov;1(3):140216–140216. Available from: http://rsos.royalsocietypublishing.org/cgi/doi/10.1098/rsos.140216.

[59] Antelis JM, Montesano L, Ramos-Murguialday A, et al. On the usage of linear regression models to reconstruct limb kinematics from low frequency {EEG} signals. PLoS One. 2013;8(4):e61976. Available from: http://dx.doi.org/10.1371/journal.pone.0061976.

[60] Müller-Putz G, Scherer R, Brunner C, et al. Better than random: {A} closer look on {BCI} results ,. International Journal of Bioelectromagnetism. 2008;10(1):52–55. Available from: http://www.ijbem.org/volume10/number1/100107.pdf.

[61] Maris E, Oostenveld R. Nonparametric statistical testing of {EEG}- and {MEG}-data. J Neurosci Methods. 2007 aug;164(1):177–190. Available from: http://dx.doi.org/10.1016/j.jneumeth.2007.03.024.

[62] Platt J. Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. Science. 1964;146(3642):347–53.

[63] Bacon F. Novum organum (gw kitchin, trans.) ; 1960.

[64] Chamberlin TC. Studies for students: The method of multiple working hypotheses. Journal of Nutritional Medicine. 1992;3(2):159–165.

[65] Klein RA, Vianello M, Hasselman F, et al. Many labs 2: Investigating variation in replicability across samples and settings. Advances in Methods and Practices in Psychological Science. 2018;1(4):443–490.

[66] Foster ED, Deardorff A. Open science framework (osf). Journal of the Medical Library Association: JMLA. 2017;105(2):203.

[67] Pernet C, Garrido M, Gramfort A, et al. Best practices in data analysis and sharing in neuroimaging using MEEG. 2018;.

[68] Enriquez-Geppert S, Huster RJ, Herrmann CS. EEG-neurofeedback as a tool to modulate cognition and behavior: a review tutorial. Frontiers in human neuroscience. 2017;11:51.

[69] Ros T, Enriquez-Geppert S, Zotev V, et al. Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). PsyRXiv, https://doiorg/1031234/osfio/nyx84. 2019;.