In this document we extensively comment on the decision of the Commission on the Investigation of Scientific Misconduct of Tübingen University. We believe that the commission relied on data that were wrongly transmitted and did not consider additional data transmitted by us and misunderstood other aspects of our research. A communication about these issues was not sought. We also would like to point out that the commission did not take into consideration that the article and comment and the data upload in PLoS Biology were reviewed by 5 reviewers in total and by the PLoS Biology Editors. Moreover, the fact that our data were replicated and published by an independent expert for machine learning, Dr. Sudhir Pathak, was not considered by the commission.

University of Tübingen
Commission for the Investigation of Misconduct in Science

Based on a report of the contact persons of the Medical Faculty (xxx) of November 22, 2018 the commission opened proceedings against the concerned persons Prof. Birbaumer and Dr. Chaudhary in its session of January 23, 2019.

The persons involved in these proceedings (§ 11 VerfahrensO) were the ombudspersons xxx, xxx (until they left their function in February 2019), xxx, xxx and xxx.

xx left the commission on April 1, 2019 due to his sabbatical. He was replaced by xxx. In its session of April 17, 2019 xxx was named as expert by the commssion (§ 13 VerfahrensO).

We are currently obtaining legal counsel to what extent the composition of the commission and the choice of expert were correct and to what extent norms relating to good scientific practice were followed and will therefore not comment extensively on this section.

But note: xxx was named expert after he left the commission and is at the same institute as the whistle blower.

On February 25, 2019 the concerned Prof. Birbaumer and on March 6, 2019 the concerned Dr. Chaudhary were heard based on § 12 S. 1 1. clause of the Code of procedure.
The informant was questioned in a session of February 27, 2019, the head of the commission also made a telephone call with him on April 2, 2019.

Both concerned renounced to be heard based on § 12 S. 1 2. clause of the code of procedure (both based on Email of May 23, 2019).

It is correct that we did not use the possibility for a hearing in front of the commission. We did not have legal counsel at that time and were in error about the legal situation. We misinterpreted that note by xxx (head of committee) that the hearing was „a necessary step as foreseen by the rules of procedure“ and that „the commission had no questions“ possibly erroneously in a way that there was no interest of the commission to hear our statement.

Xxx (head of commission) also wrote that the commission had downloaded the data we had sent to the DFG on 17.05.2019 and had discussed them. We

have confirmed that download of 20.05.2019. Since these excel files contained comprehensive information about the data, we are unsure if the commission could really appreciate their content without relating them to an expert and assumed that the commission was not really interested in thoroughly investigating the problem. Given the complexity of the issue, we were also not convinced that the commission was truly taking our responses and arguments properly into account. We had the impression that the commission acted under great time pressure since a proper evaluation by an external expert should take months in such a complex matter but it was obviously performed in 4 weeks. For example, the review process took almost a year to complete. The data collection and analysis and the writing of the article took 3 years altogether. This procedure strengthened our impression that the commission had not really dealt in detail with this very complex matter. We therefore did not see any sense in talking again to the commission especially since its head, xxx, wrote: "To avoid misunderstanding: The committee does not have further questions or the like. The hearing is optional." After obtaining legal advice we do now understand that this was a mistake.

In its session of May 25, 2019 the commission came to the following unanimous decision:

Decision

The authors Prof. Birbaumer and Dr. Chaudhary have committed scientific misconduct
.

I. Issue

The concerned published together the article „Brain-Computer Interface-Based Communication in the Completely Locked-In State" in the Online-Journal PLOS Biology" (ISSN 1544-9173; eISSN 1545-7885) (DOI:10.1371/journal.pbio.1002593; Day of publication January 31, 2017).

This publication is based on data that were recorded during several sessions with four patients, who were in an advanced state of amyotrophic lateral sclerosis. These patients are no longer able to communicate with their environment based on the total loss of movement control, even of eye and eye lid movements. Their state is therefore called „„Completely Locked-In State" (CLIS).

Here the method of functional near infrared spectroscopy (fNIRS) was used. Using techniques of machine learning, an attempt was made to develop a model that permits by applying it to a set of data to achieve a statistically valid assignment of the brain activity of a patient to a yes or no answer of the patient based on his or her thinking of yes and no. In this sense this is called a brain computer interface (BCI).

In the publication the authors claim that this setup permits to assign the brain activity of the patients with a clearly above chance probability to a yes or no answer (that is thought by the patient (article: p. 1 „Online fNIRS classification of personal questions with known answers and open questions ... resulted in an above-chance-level correct response rate

over 70 %."). We refer to the Report, pages 2-3.

The research group that was led by Prof. Birbaumer made the data available to the later whistle blower Dr. Spüler. The whistle blower is himself active in the area of BCI and cooperated for several years with Prof. Birbaumer in this area. We refer to the Report, page 4.

Dr. Spüler raised doubts on the capability of the method the authors presented in a „Formal Comment". This text was originally rejected by PLoS Biology based on the evaluation of two reviewers in a review process of December 2017 and sent back for revision, the revised text was then rejected in March 2018 (see report, attachment 2B). When Dr. Spüler intervened against this decision, his revised text was published on April 4, 2019. (https://doi.org/10.1371/journal.pbio.2004750) (see Report, attachments 3 and 4). At the same time the „Response to: ‚Questioning the evidence for BCI-based communication in the complete locked-in state'" by Prof. Birbaumer and Dr. Chaudhary was published (https://doi.org/10.1371/journal.pbio.3000063).

After the hearing on 06.03. 2019 the commission asked Dr. Chaudhary to provide all data that are needed to replicate all steps that were made in the article as they had been presented by him (Email of 28.03. 2019). Dr. Chaudhary has subsequently provided data in the magnitude of about 5.2 GB. These are the NIRS data of the four patients F, G, B and W, that are subdivided in individual folders „visit [n]" and subfolders that are marked with a date (e.g., „2014-06-10") and two scripts that were used in the data analysis.

Prof. Birbaumer sent to the head of the commission an Email of May 18, 2019, that contained the letter to the DFG of the concerned persons of May 17, 2019 and a link where further data were presented for download (that had to be authorized by the concerned persons) („Additional Data for DFG", containing the archive „17052019_Reply.zip"). In the letter and the data that were provided, the concerned admit that there are missing data, that data were excluded from the data analysis and that data that were collected during the experimental sessions, were later renamed. The concerned also describe that some training sessions were text validation sessions that are not described in the article (p. 17).

We strongly argue against this interpretation of our letter to the DFG. We had the impression that some of the questions raised by the whistleblower resulted from a misunderstanding of the description of the data and procedures, which was not intended by us. We therefore made a major effort to describe the procedure and what exactly was done in a much more detailed manner than requested by the journal. This additional information should now not be held against us, especially since it does not change anything at all in the algorithm we used to assess communication ability in the patients. It only documents how we handled problems arising in the interaction with the patient in a very detailed manner and spells out what happened in every single session. The additional information also presents in detail the inclusion and exclusion of training trials that were used to arrive at the classifier. It also documents the feedback sessions, which were used to determine communication ability. Whereas the patients did not get feedback if the answer was right or wrong in the training sessions, which were used to build the classifier, the patients received feedback if their answer was right or wrong in the feedback sessions, which were used to determine if the patients could communicate. We included all session to determine communication ability in the analysis even if the patients could not differentiate yes/no. This was the case in 7 of the 21 sessions that

were analyzed for communication ability and the total percentage of correct answers was also based on these sessions.

The sessions we called text validation sessions were normal training sessions and treated as all the other training sessions. There was therefore no need to mention them separately in the PLoS Biology paper of 2017. We only mentioned these sessions because we think that it is possible that the whistleblower read their name in the data we shared with him and erroneously viewed them as feedback sessions, which were used to determine the percentage correct responses to assess communication ability. However, the text validation sessions only had the feature that the experimenters could see the result of the prediction of the answer and thus see if the predicted answer from the brain and known answer of the question matched. These are not feedback sessions and were not used to calculate yes/no percentages and thus the probable communication ability of the patients. As noted, the independent recalculation of the data by Dr. Sudhir Pathak yielded even higher estimates of communication ability.

We would like to emphasize that the data that were excluded due to equipment failure are never reported in any paper we know and they could not have been analyzed as the triggers could not be properly assigned.

Data were never renamed but only the name of the sentence in the text file that was used for the program. Sentences with open questions are prefaced by 003 and the correct answer to the sentence is always indicated by 2 since it is not known what the right answer would be. In contrast, text files for sentences for the feedback questions are prefaced by 001 and 002 depending on if the sentence represents a true or false statement and then the correct answers to the sentences are labelled 1 and 0. Thus in this specific session open questions were used but presented in the file for feedback questions. Then the normal procedure for calculating the percentage accuracy for open questions was used.

No data content was altered. This was a totally acceptable procedure, since this is only a formal operation. To prove this we also provided a folder with a time stamp that shows the renaming and the original folder. There were always 2-3 persons with the patient and they can witness this procedure. Moreover, open question sessions were not used to document the patients' communication ability – they were only reported to show how they were used and to encourage the readers to consider using them with patients to determine their needs. So, this has no consequence for the feedback sessions used for determining if the patients could communicate.

II. Standards used for the assessment of scientific misconduct

The term of scientific misconduct in science is described in § 1 VerfahrensO . In § 1 Abs. 2 individual behaviors are named pertaining to three groups of cases (false statements, infringement of intellectual property, impairment of the research activity of others).

In contrast to the rules of procedure („Code of procedure") in dealing with scientific misconduct of the DFG, the rules of procedure of the University of Tübingen do not deal with these three groups of cases but defines in in § 1 Abs. 1 misconduct as „behavior in a science-relevant context that that violates legal statutory provisions or written or unwritten rules, the adherence to which is viewed as indispensable in a certain scientific field or a

scientific institution." The groups of cases of § 1 Abs. 2 are thus, as the formulation „in particular" implicates do not refer to an exhaustive list.

Thus the commission has not limited its examination on the groups of cases in § 1 Abs. 2 but also examined if rules were violated in the case at hand the adherence to which is viewed as indispensable.

On the other hand, regulation § 1 Abs. 1 and 2 of the Code of procedure, implies that not every scientific mistake is scientific misconduct. Scientific misconduct is a case of dishonesty not of error.[1] Erroneously methodologically wrongly designed experiments, errors in thinking, erroneously wrong or omitted use of relevant statistical methods etc. as well as unclear, undetermined or contradictory statements are as such not scientific misconduct.

Thus quality assurance in a certain area of research is the task of the respective scientific community; the commission for scientific misconduct is not called upon this task and is not allowed to place itself at such a disposal. The question if such shortcomings are present cannot be the topic of the examination of such a commission.

Therefore, one of the facts and circumstances named on page 5 of the report in a listing with five subsections (implausible statements in the publication) was not considered from the beginning of this procedure.

III. Existence of misconduct

1. Selective choice of data in data acquisition

a) In the article on page 17 the following is written (i.e. in the description of the experiment):

> „Three to four sessions were performed each day depending upon the health condition reported by the caretakers of the patient. Every sessions lasted for 9 min, and a session in progress was termi- nated extremely rarely (i.e., if removal of saliva became urgent). In such a rare event, the session was started again. … A session, once in progress, was never terminated for patients F, G, and W. For patient B, a session was terminated while in progress three times because of removal of saliva, and the data were not included in any kind of analysis. … Each BCI session started with training ses- sions, ..."

---

1 Compare Memorandum of the German Research Foundation („Recommendations for the Assurance of Good Scientific Practice" (1998, amended 2013), p. 13 and 40. On page 9 of the article it is stated: „None of the sessions were eliminated in the analysis, and only very few sessions had to be interrupted because of live-saving measures such as sucking saliva; thus, no bias for selecting 'successful' sessions incriminates the results."

b) In his email of 9/10/2017 Prof. Birbaumer states the contrary: „certainly we eliminated some sessions when family and patients werent fit, thus biasing the results toward positi- ve." (Report, attachment 1B).

The email of Professor Birbaumer referred to exactly the situations we mentioned in the paper (p. 17, lines 23-25): "A session, once in progress, was never terminated for patients F, G, and W. For patient B, a session was terminated while in progress three times because of removal of saliva, and the data were not included in any kind of analysis."

There we indicated that we had to exclude sessions where the patients could not participate in the experiment due to immediate health care issues (e.g., threat of suffocation, elimination of saliva) based on the information of family members and care takers. This does not refer to any state of the family member.

In the email v. 16/10/2017 Prof. Birbaumer writes furthermore: „… Ujwal and I did most of the experiments the last years together and I pressed him often to eliminate a session if the patient state requested that, ..." (Report, Attachment 1H, p. 1).

In the hearing of 25.02.2019 Prof. Birbaumer said that such an exclusion wmade "sometimes", Asked for the criteria for such an exclusion he pointed that he personally could judge these situations.

c) In the letter to the German Research Foundation of 17.05.2019 , which was sent to the head of this commission by Prof. Birbaumer per email on 18.05.2019 (under 4. a.E.) the concerned persons say the following: „This means that we … excluded data in the model building stage when the state of the patient did not permit differentiation of yes, no states ..." This shows that „sessions" were excluded in cases other that those related tot he health of the patient based on an (unclear) criterion "the state of the patient did not permit the differentiation of yes and no answers.

We cannot follow this judgement. We clearly stated in the paper on page 19, lines 3-6, that we included only sessions that exceeded chance differentiation based on the NIRS signal. ("If the classification accuracies for at least three consecutive "training sessions" with questions with known answers were greater than the chance-level threshold, a new model was generated using the relative change in O2Hb across three training sessions to give online feedback."). See also "BCI Effectiveness Metric" on page 18, where we describe how chance level was defined.

We would like to reiterate that these are not laboratory experiments but assessments at the patient's home with many types of difficulties that have to be taken into account when the data are acquired because, otherwise, invalid data would be collected. Since the commission and the expert never attended the data collection in the patients' home, we believe that it was very difficult for the commission to correctly evaluate the data collection and analysis process. The entire experiment was part of the Koselleck project of the accused Birbaumer, where risky experiments that go to the limits of what is possible, are encouraged.

d) In the letter the concerned persons further admit (under 4.) that the data acquisition had technical problems that often led to the exclusion of "sessions". "There were many instances when there was an error in the online data transfer ..."). The concerned persons state: „We have marked these files in the attached excel files as ‚Data from this session was not analysed because of an online data transfer problem'. The

data were thus acquired and saved but were not processed because of this error.“ The file that was provided for download by the concerned persons „Readme_SessionDetails.docx“ contains a list of the excluded „sessions“. These were 10 in F, 7 in patient G, 2 in patient B and 1 in patient W, 20 total. This is in direct contradiction to the statement in the arcticle, p. 9: „None of the sessions were eliminated in the analysis, ...“

These sessions, which were given to the DFG for examination, are sessions where hardware/software interaction problems led to data that could not be analyzed and the session had to be repeated. This was related to trigger problems that made it impossible to analyze the data, as can be seen in the data. At no point were unwanted data excluded due to not correctly transmitted triggers. The trigger problems made it demonstrably impossible to analyze the data. We thus have at no time excluded unwanted data, since the error in the machine-based data transfer made it impossible to analyze the data.

This can be verified at any time by the additional data provided to the DFG, which were also available to the committee. Since we considered this an equipment failure, we did not report the sessions as they would likewise not be reported in other experiments and because no data were collected that could be evaluated. For example, in magnetic resonance imaging, sessions with trigger failure are simply repeated after the failure has been fixed. This is exactly how we handled these sessions. If the committee doubts this procedure and we wrongly did not upload these data, we will of course add these sessions to the uploaded data base. But we would again like to reiterate that PLoS Biology only requests upload of data that pertain to the results. Since these data could not be analyzed and were not included in the results, there was in our judgement no need to upload them.

This question was also answered in the letter to the DFG dated 17 May 2019 as follows:

> “All the data of the feedback/open question sessions were included in the data analysis and thus the publication. We only excluded data in the model building stage (training sessions). The model was built using training sessions where the differentiation between the yes and no exceeded 65% (see page 19, lines 3-7 of the paper) as described in the attached excel file of each patient. In 2014, during the experiment there was online data transfer provision between the data acquisition and BCI software laptop. There were many instances when there was an error in the online data transfer between laptops leading to the loss of data packets and hence to the loss of trigger markers. We have marked these files in the attached excel files as “Data from this session was not analysed because of an online data transfer problem”. The data were thus acquired and saved but were not processed because of this error. This means that we included all data in the feedback/open question session and only excluded data in the model building stage when the state of the patient did not permit differentiation of yes, no states or the trigger was not functioning properly and thus precluded model building. This is indicated in the readme files we included.”

And

> “In Table 1 of the original publication we included the total number of training sessions, feedback and open question trials per patient. In the new excel sheet which we are including with the data we also listed the training trials which had data transfer problem as described above (these were not included in the PLoS Biology publication

since they were not analysed due to online data transfer error as described above in point 4 and were not uploaded for PLoS Biology but were uploaded for the DFG). The same holds true for the figures."

And

"Table 1 gives the total number of sessions that were analysed. Here we included all the sessions that were run i.e. also those with below chance level but not the sessions with wrong triggers. The sessions with wrong triggers were only provided to DFG since they were related to equipment problems, but we wanted to be complete in the provision of even excluded data."

e) The concerned persons excluded „sessions" due to

- The health state of the patient (often)
- Technical problems(often
- „State of the patient does not permit differentiation of yes and no answers"

An exclusion of certain „sessions" and the data that were acquired therein is as such not inadmissible. However, the criteria for this exclusion must be defined and documented before the beginning of the data acquisition. In addition, the readers of the article must be informed about the number of "eliminated" sessions as well as about the criteria used and the decision process about the elimination. This has not happened. Neither in the article p 8 („Slow EEG Rythms' Relati- onship with fNIRS Classification Accuracy") nor in the „Response to: ‚Questioning ...'", p. 3-4 („Slowing of EEG and consciousness") any mention is made how many sessions were excluded based on which criteria. Rather, in the article on page 9 there is a conscious false statement that no sessions were eliminated from the analysis, especially the concerned persons also fail to mention the technical problems. The NIRS data that were given to the commission also provide no information which „sessions" were excluded based on which criteria. The ex post created tables of May 2019 about the exclusion of data based on certain factors, that name, moreover, exclusion based on „state of the patient did not permit differentiation of yes, no states" without any further explanations, does not change this.

The described procedure is thus scientific misconduct based on § 1 Abs. 2 Nr. 1. b) (Falsification of data by refuting unwanted results without disclosure).

We cannot follow this argument. As stated in the paper (page 17, lines 23-25), "A session, once in progress, was never terminated for patients F, G, and W. For patient B, a session was terminated while in progress three times because of removal of saliva, and the data were not included in any kind of analysis." The health state of the patients required this exclusion of sessions in the model building stage only sometimes.

The uncertainty of the patients' state as determined by the lack of yes/no differentiation required the exclusion of sessions more often as stated on page 19, lines 3-6 ("If the classification accuracies for at least three consecutive "training sessions" with questions with known answers were greater than the chance-level threshold, a new model was generated using the relative change in O2Hb across three training sessions to give online feedback.").

Moreover, technical problems also occurred often. We did not describe the equipment malfunction sessions, which were repeated, and their report is in our opinion not necessary since they could not be analyzed and are not either reported in the literature. We documented all the remaining decisions in detail also in the paper and data exclusion based on nondifferentiation of yes/no refers to the training phase which served as the basis of model building. In other words, we excluded sessions to be able to build a model for yes-no answers in these very disabled patients with compromised brain activation patterns.

We never excluded sessions when we gave feedback to the patients, which served as our determination of communication ability. Given the very difficult conditions, we believe that we tried our best to obtain a valid model and a sound basis for an algorithm designed to later assess the patients' communication ability. In the feedback sessions, where we judged communication ability, we never excluded sessions even if we thought that the patient was sleeping or otherwise unwell in order not to bias the data in our favour.

To reiterate: after a lengthy training phase, where we excluded sessions from the analysis to be able to build a mathematical model to differentiate yes/no indicating brain states, we accepted those sessions that permitted above chance yes/no differentiation and built a classifier in a known and accepted fivefold cross-validation procedure. We then used the thus built classifier to judge communication ability in a new set of sessions, the so-called feedback sessions, where we gave feedback to the patients and where we did not exclude any sessions.

In addition, we have to state that Prof. Birbaumer was well aware that his behavior had led to a statistically relevant distortion („bias") in the data whereas in the article the opposite is claimed (see above). In addition, in the article p. 17 the interruption of „sessions" is characterized as „rare" event, whereas Prof. Birbaumer claims in relation to the years where also the "sessions" that are at stake here took place writes that he had often urged to exclude a "session" („I pressed him often to eliminate a session", see above.).

Again the commission misunderstands what we did in the training sessions for the model building and the feedback sessions that were used to determine communication ability. Training sessions served to build the mathematical model whereas feedback sessions were the basis for the determination of the communication ability. In the training phase we only used sessions where the patients could clearly differentiate yes/no and this is certainly a bias in the sense that we tried to maximize the minimal communication ability these patients may have in optimizing the algorithm. When in doubt, the session was excluded rather than included because we have no objective means to determine the state of the patient. This was always based on the NIRS differentiation of yes/no answers.

This is different from interrupted sessions due to the health state of the patient, which were rare and did of course not enter the data analysis stage at all. But as stated before, we biased that data rather against us in the feedback phase, which was used to determine communication ability and where we never excluded a session even if we thought that the patient was not fit. This is also what page 9, lines 5-7 of the discussion, refers to where we wrote „"None of the sessions were eliminated in the analysis, and only very few sessions had to be interrupted

because of life-saving measures such as sucking saliva; thus, no bias for selecting "successful" sessions incriminates the results."

Finally, the following statement suggests that Prof. Birbaumer attributed the ability to himself to differentiate yes-no answers during a running session with the patients on site: „I am positively biased like all, but not to the extreme that I cannot differentiate yes from no and that in almost 100 sessions with several patients " (Email of Prof. Birbaumer of 16/11/2017, Report, Attachment 1L, p. 1).

This is true in the sense that an experienced fNIRS person could see if there was differentiation of the data in yes/no based on the brain signal. As noted in the paper, the exclusion of sessions was always based on the documented lack of yes/no differentiation in the machine learning algorithm that was used and not on the communication between the experimenters during training. The commission clearly cites a statement of Professor Birbaumer out of context. Professor Birbaumer discussed in this context the role of the physiological signal versus the machine learning algorithm and the theoretical basis of the paper, since there is a discussion if a machine learning algorithm alone can really capture the nuances of the physiological signal. However, this is a more general problem and in the PLoS Biology paper the machine learning algorithm was used as described there.

f) Additional evidence

The email of 16/10/2017 of Prof. Birbaumer (Report, Attachment , Anlage 1G, p. 1) contains the following statement:

> „Right now we have to wait for Ayala to send us back the data, … In his case I was present during most sessions and I judged the performance [sic] by deciding visually also whats no and whats yes according to the shape of the physiological signal. That correlated perfect with the classification but we eliminated all trials where it did not correspond to my judgement of the physiological signal. That may introduce a bias but its better than blind model building in these high variance data."

Here again the initial appearance is given that personal decisions were used without further criteria to eliminate data („we elimina-ted ... trials"). The described circumstance is, however, mentioned in connection with the publication *Gallegos-Ayala/Furdea/Takano/Ruf/Flor/Birbaumer*: Brain communication in a completely locked-in patient using bedside near-infrared spectroscopy, in: Neurology 82 (2014), S. 1930-1932.It was, therefore, not part of the current investigation.

This is cited out of context and does not even relate to the PLoS Biology paper, because we had intensive discussions on the scientific hypothesis of the experiment, i.e. how the shape of the physiological signal correlated with the classification. As noted above, when in disagreement, the model was used, not the visual inspection. This discussion focussed on the question if physiological data and the classification by the model should always correspond. As noted in the paper, visual inspection was not a criterion for including training sessions in the model. We object to the commission using quotes from scientific discussions unrelated to the paper out of context.

In reference to this article it has to be stated that the data that were used there may have been compromised by a wrong handling of the fNIRS equipment, as written in the email of. 20/10/2017 (Report, Attachment 1J, p. 1f.).

Prof. Birbaumer has confirmed the possibility of such a mistake in an email of 20/10/2017 („The trigger problem … it is only relevant for the Ayala et al. paper … Guillermo [Gallegos-Ayala] … he has the calender [sic] of the Hitachi use."; Report, Attachment 1K, p. 1).

The commission ignored our response on this issue to the DFG of April 22, 2019, which we also made available to the commission. There we wrote: "We did not find any fault in the 2014 Gallegos-Alaya paper and do not believe that the whistleblower has provided any proof of scientific misconduct. The data of the subject where the whistleblower mentioned a wrong trigger were never included in this report since this was not a patient but a pilot healthy subject. The whistleblower had no access to the data of Gallegos-Ayala, any claims he made or makes about these data are not based on real data but assumptions and accusations. In fact, the trigger type he mentioned is not used in the NIRS device used in the 2014 paper (see Email by Gallegos-Alaya, attachment 6, evident also in the user manual of the device, and we are afraid that the accusations are not based on scientific evidence). The fNIRS equipment used in the PLoS Biology paper is not the same as that used in the Gallegos et al. 2014 paper but a more advanced type, since all data of the PLoS Biology paper were collected from 2014 on. There were no trigger problems of this type for the data in the PLoS Biology paper. The statements of the whistleblower referred to other days and other experiments and a different fNIRS machine. Thus, none of these statements is valid.

2. Missing disclosure of data and scripts

a) The article contains links to various data sets. Among them the script which was used for the analysis program that describes the rules for data use.

In addition, the article does not contain links to the data that would permit to determine how the described model would provide statistically valid assignments of the respective patterns of brain activity of the patients to a yes no answer thought by the patient. Links are present („S4 Table – S11 Table"), which relate to the respective training and feedback sessions of patients F, B, G und W. These only point to Tables (in MATLAB and Excel format), that contain nothing but the data of the graphics of the article, not the data themselves, i.e. the final result.

b) aa) Thus the publication does not adhere to the guidelines of the journal, in which they were published. They are contained in a „Data Availability Policy", which is valid for all PLOS journals. (https://journals.plos.org/plosbiology/s/data-availability). There it is stated:

> „PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction at the time of publication. When specific legal or ethical requirements prohibit public sharing of a dataset, authors must indicate how researchers may obtain access to the data.

> When submitting a manuscript, authors must provide a Data Availability Statement describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the accepted article.
>
> Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.“

bb) In addition, the publication goes against recommendation 7 of the commission „Self control in science“, that is contained in the memorandum „recommendations fort he Assurance of Good Scientific Practice” (1998, amended 2013), there (p. 21f.) is stated:

> „Experiments and numerical computations can only be reproduced if all important steps can be reproduced. For this they have to be documented. Each publication, no matter if based on experiments or numerical simulation, contains an obligatory section ‚Materials and Methods‘, which summarizes the documentation in a way that the work can be replicated elsewhere..“

c) In the hearing the concerned persons have claimed that they were not asked by the editor to provide additional data. Dr. Chaudary has affirmed that he would provide the raw data on request to other scientists. He has provided data of the scope of about 5.2 GB tot he commission upon request. These are the NIRS data of the patients F, G, B und W, which are subdivided in folders „visit [n]“ and subfolders with a data (e.g., „2014-06-10“) as well as scripts that were used for data analysis (compare o. I. a. E.). On the basis of these data the statements of the article cannot be reconstructed because the information is missing which questions were used and what the BCI system had computed. In addition, the EEG data that are mentioned in the article are not contained.

With the email of 18/05/2019 Prof. Birbaumer provided a Download-Link (cf.. o. I. a. E.) by which the data can be accessed and by which in principle for a part of the experiments the NIRS data can be linked to the sound files and the text files. In the Table „SessionDetails“ the „Folder Identifier“ refers to the NIRS-data and permits on the other hand with the assignment of „Session“ to relate the data to the „Feedback results“ (for „feedback sessions“, for patient B for example 4) and „Open question results“ (for „open question sessions“, for patient B for example 2). These „results“ are two tables of which one refers tot he sound files (die questions). The other contains the answers „for the „feedback-sessions“ that were created by the BCI-system, and for the „open question sessions“ the estimate of the correctness of the computed answers (but see on that 4 below.).

Based on the statements of the concerned persons these are new summary files that were created at the time the letter was sent to the German Research Foundation. („we have now also added“; „new excel sheet“, letter of 17/05/2014 under 2. und 5., see also the date of all files). We thus must assume that they were non-existent at the time when the NIRS-data were given to the commission.

Independent of this there are still missing data on the “training sessions”. Thus the reconstruction and the testing of the model development and selection during the training sessions is still not possible. The data set is thus still incomplete. Finally, the EEG data that are described in the article, are missing.

The concerned persons have not delivered complete data neither with the publication nor to the commission at the end of March nor to the DFG mid May.

We strongly disagree with this statement and would like to emphasize that we uploaded all data to PLoS Biology as requested by the reviewers and editors and the PLoS Biology data policy, which by no means requires that ALL data be provided as the quote from the data policy provided further down shows. In addition, to help the committees with their evaluation, we provided additional detailed information on the data acquisition and analysis in a much more extensive manner than usually required by journals such as PLoS Biology. This effort to aid the committees of the University and the DFG in their evaluation should not be held against us. Also the fact that we wrote "we have now added" should not be held against us since we compiled additional information for the reviewers of the DFG and university commissions, which were available but were never asked to be provided by the PLoS Biology reviewers and editors who also reviewed our uploaded data.

Data policy of PLoS Biology:

"DATA POLICY:
You may be aware of the PLOS Data Policy, which requires that all data be made available without restriction: http://journals.plos.org/plosbiology/s/data-availability. For more information, please also see this editorial: http://dx.doi.org/10.1371/journal.pbio.1001797

Note that we do not require all raw data. Rather, we ask that all individual quantitative observations that underlie the data summarized in the figures and results of your paper be made available in one of the following forms:

1) Supplementary files (e.g., excel). Please ensure that all data files are uploaded as 'Supporting Information' and are invariably referred to (in the manuscript, figure legends, and the Description field when uploading your files) using the following format verbatim: S1 Data, S2 Data, etc. Multiple panels of a single or even several figures can be included as multiple sheets in one excel file that is saved using exactly the following convention: S1_Data.xlsx (using an underscore).

2) Deposition in a publicly available repository. Please also provide the accession code or a reviewer link **so that we may view your data before publication.**

Regardless of the method selected, please ensure that you provide the individual numerical values that underlie the summary data displayed in the following figure panels: (e.g. Figs. ....), as they are essential for readers to assess your analysis and to reproduce it. Please also ensure that figure legends in your manuscript include information on where the underlying data can be found.

Please ensure that your Data Statement in the submission system accurately describes where your data can be found."

We would like to add that we have clearly defined which training sessions were excluded in the paper and the expert could easily have found which sessions were excluded as this was explicitly stated in the paper (a new model was only built when 3 sessions in a row were above chance).

d) The rules of procedure of Tübingen University do not contain a provision that matches the above mentioned recommendation 7 of the commission “self control in science”. However, the duty to provide raw data and scripts to enable other scientists verification, could constitute an indispensable written rule within the meaning of § 1 Abs. 1 of the rules of procedure.

aa) The rules of procedure refer in this context to the „written or unwritten rules……in a certain scientific specialty or a scientific area“, i.e. the disciplinary culture. The commission cannot determine that the concerned scientific area views it as indispensable that it is necessary to provide all raw data and scripts together with the publication. In some instances in medical research an immediate and complete publication may even be impermissible.

bb) From this the question if an incomplete provision of data to the commission is in itself scientific misconduct has to be separated. The commission is aware that it has a special role as a body that was instituted to clarify the allegation of scientific misconduct: as examining body it is in a conflict in the role of a decision making body when there is the allegation of scientific misconduct towards the concerned persons. It must therefore be considered that the behavior of those that are concerned does not turn into scientific misconduct in the course of an examination by the commission which it would not be outside the realm of the investigation.

In the present case Dr, Chaudhary has explicitly stated that he and Professor Birbaumer would give full insight into the data to all but the whistle blower. Dr. Chaudhary affirmed several times that they wanted to create total transparency. Thus the commission did not ask for more than the concerned persons offered by themselves..

The incomplete transfer of the relevant and for the reproducibility of the experiments described in the article necessary data to the commission is scientific misconduct based on § 1 Abs. 2 Nr. 1 a) (falsification of data by suppression of relevant records)..

We do not agree with the evaluation of the commission. We provided the in our opinion best available description of data and methods when we uploaded them to PloS Biology and strictly followed the PLoS Biology data policy. We did not upload the EEG data because they were not the topic of the paper and were only mentioned as non-sufficient to allow communication in these patients. The EEG data were never asked to be uploaded by either the reviewers or the editors of PLoS Biology. They were also not part of the results related to the fNIRS communication ability.

We also shared all data relevant for the paper with the commission. However, in the course of the questioning by the committee members of the DFG, that were very detailed and clear, we realized that more information might be useful to fully understand each detail of this publication and we provided much more detailed excel sheets for this purpose. Overall, we spent more than 2 years on this very complex analysis process (and 1 year on data collection) and we tried to give a concise and still comprehensive description of the data and analysis process. We also declared that we would be happy to answer all questions on this process to anyone. For us as scientists deeply involved in the analysis of BCI data and the related procedures, it is not easy to judge how much documentation and explanation

persons outside this field would need to understand the data collection, analysis and replication process. We learned from this process that even more detailed documentation might be desirable in future publications and can also provide this for the publication under scrutiny. Again, none of this was done on purpose or to preclude replication. On the contrary, we would be most interested in replication. In fact, Dr. Sudhir Pathak replicated our data and published the results with us.

As noted above, the overall 8 reviewers of our two publications (original and commentary) and the editors of PLoS Biology found the description of the data analysis and the provided data sufficiently detailed. The data provided in PLoS Biology were thus reviewed by the reviewers and the journal editors who were happy with the results and the data and they never asked that we should make any further modifications. Moreover, the data policy of PLoS Biology does NOT require to publish all data but only those summarized data that are relevant for the figures and results as evident from their data policy we have included in the last section. We would like to point out that PLoS Biology does not require complete upload of all raw data but only of those data that support the conclusions.

Moreover, we believe that the commission was in possession of all data needed to reproduce the results. It is possible that the expert who was called in, did not understand every aspect of the documentation and analysis. We would have been happy to provide additional information had the expert or commission requested this from us.

3. Missing data

a) NIRS-data: missing days with sessions (patient B)

In the article information is provided on the days when the patients participated in „training or feedback and open question sessions“ (article, p. 11, Table 2 with Fn. 3). From this Patient F underwent 14 days, patient G 17 days, patient B 12 days and patient W 6 days of sessions with NIRS measurement. This is matches the graphic displays on pages 7-10, where the results of 14, 17, 12, and 6 days of the mentioned patients are presented (see also the article page 20:
„The number of days for each patient were: F, 14; G, 17; B, 12; and W, 6.“). In the article p. 18 one can read: „The fNIRS data was acquired online throughout all the sessions, namely trai- ning, online feedback, and open questions sessions.“

In the transmitted data (cf. o. 2. c)) are folders with results of the following days (left column), here juxtaposed with the sessions that were named in the article (right column)

Patient F			days mentioned in the article
	visit 1	24.-28.03.2014: 5 Days	
	visit 2	15.-20.05.2014: 6 Days	
	visit 3	04.-07.08.2014: 4 Days	
	visit 4	04.-07.11.2014: 4 Days	
		Sum: **19** Days	*14*
Patient G			

	visit 1	17.-21.06.2014: 5 Days	
	visit 2	25.-31.08.2014: 7 Days	
	visit 3	21.-26.09.2014: 6 Days	
		Sum: **18** Days	*17*
Patient B			
	visit 1	10.-12.06.2014: 3 Days	
	visit 2	12.-16.08.2014: 5 Days	
		Sum: **8** Days	*12*
Patient W			
	visit 1	03.-08.09.2014: 6 Days	
	visit 2	15.-19.12.2014: 5 Days	
		Sum: **11** Days	*6*

In no case the number of days for which data are resented is in accordance with the number of days for which the article presents results. Thus the statement of the article page 11, table 2, Fn 3 is wrong. Patients F, G, and W had more sessions that were not entered in the analysis. It cannot be reconstructed why these sessions were excluded.

Specifically, for patient B 12 days are indicated, however, there are only data for 8 days. Thus data are presented for days where no data exist.

This is scientific misconduct based on § 1 Abs. 2 Nr. 1 a) (Invention of data).

We never saw the data that are presented in the left-hand column. When we examined how those data might have come about, we checked all stages of our data documentation and data transfer. We noticed that there was a data transfer problem in our submission of data to the University committee on the 2nd of April that resulted in some extra data being transferred and few others being omitted. Specifically, Patient F - Visit 4 (2014-11-04 to 2014-11-07) - later session, which was not included in the publication but erroneously given to the committee; Patient B - Visit 3 (2014-08-04 to 2014-08-07) included in the publication but erroneously not provided to the committee; Patient W - Visit 2 (2014-12-15 to 2014-12-19) – later session, which was not included in publication but erroneously given to the committee.

We did not notice this error at that time. We provided the correct number and sequence of sessions in the excel sheet mentioned in our letter to the DFG dated May 18th, 2019. In the letter of the commission on May 22, Dr. Forster stated that the new data information had been downloaded and discussed. At this point the committee should have noticed that the data from April 2 and May 18 did not correspond. Even before that, the committee noticed and held against us that the data transmitted on April 2 did not match those of the paper. We apologize for this oversight, but we noticed ourselves only now that the transmitted data did not match the uploaded data. Dr. Chaudhary believes that there had been an oversight in compiling the data for data transfer but the exact cause for the partially wrong transmission cannot be reconstructed. Nonetheless, we would have expected the committee or the expert to ask us about the discrepancy, which could have been detected and resolved. The committee downloaded the relevant excel sheets on May 20 and wrote to us on May 22 that there were no further questions. We assume that the committee overlooked this discrepancy in the data.

Furthermore, this has as a consequence that the time line of the experiments in the article, page 17 („Each patient was visited 4 to 5 d in a month, except patient W.“) is not correct. Patients F, G und B were visited between 3 und 7 days per month.

This is true, it should have said “on average”. We apologize for the error.

b) NIRS-data: missing subfolders (patient F)

The NIRS data that were given to the commission (cf. o. 2. c)) fort the four patients F, G, B and W are organized into individual folders „visit [n]“ and subfolders that are marked with a date (e.g., „2014-06-10“). In these subfolders there are additional ones with a dat and a number, and in them again the subfolders:

„Conditions“, „Detectors“ and „nirsSPM“ (with subfolder „nirs_data“). For patient F the regular folder structure is thus (Excerpt):

Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_002
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_002\Conditions
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_002\Detectors
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_002\nirsSPM
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_002\nirsSPM\nirs_data

Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_003
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_003\Conditions
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_003\Detectors
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_003\nirsSPM
Patient F\raw\NIRS\visit1\2014-03-24\2014-03-24_003\nirsSPM\nirs_data

The respective folder (ze.g., 2014-03-24_002) contains several data sets , among them 3 with more than 3 MB. The subfolders contain smaller data sets ; the subfolder „nir- sSPM“ has only one subfile; the subfolder „nirs_data“ three data sets with about 4 MB; so for example:

NIRS-2014-03-24_003_detector_DeoxyHb.mat
NIRS-2014-03-24_003_detector_OxyHb.mat
NIRS-2014-03-24_003_detector_TotalHb.mat

This structure is, however, not consistently maintained. Thus several times the folders with running numbers (for example Patient F, 2014-03-24: present 002, 003, 004, 005, 009; not present are 001 and 006-008) are missing. Several times the folder „nirsSPM\nirs_data“ (thus for patient F, 2014-03-25_004, 2014-03-26_001 as well as 002 and 2014-03-27_001 and 002; for Patient B, 2014-06-10_001 und 006) is missing..

For patient F the folder „nirsSPM\nirs_data“ is completely missing for the days from 15.05.2014 to 20.05.2014 (6 days) as well as for 06.11.2014 (1 day). In the first time frame no subfolders are present, for the later named day only the subfolder „Detectors“.

The missing sessions were time periods where the equipment (NIRS device) was tested. This is also coded by the NIRS device even if no BCI sessions were run. As for the missing nirsSPM/nirs-data: this refers to the processed data, which were not always included but is of no consequence, since this is a later processing stage that is not needed for documentation. The same holds for the folder detectors. We are

surprised that the committee never asked us about these sessions, since this could have been easily resolved by us.

Thus within the 19 days, where sessions were run with patient F, there are NIRS data missing of seven days. There are thus only NIRS data for 12 days. In the article, data for 14 days are presented. This means that results are presented for days where no data are existing.

This difference is related to the erroneous file transmission on April 2, the true data and the data in the paper are 100% concordant and we could have solved this discrepancy had the commission asked us about it.

This constitutes scientific misconduct according to § 1 Abs. 2 Nr. 1 a) (invention of data).

We strongly disagree with this evaluation. This difference is related to the erroneous file transmission on April 2, the true data and the data in the paper are 100% concordant and we could have solved this discrepancy had the commission asked us about it. Moreover, the file the commission downloaded on May 20th, and discussed before inviting us, contained the right sessions.

Special relevance is given to this finding by the circumstance that the time period of 15.05.2014 to 20.05.2014 contained – based on the overview of May 2019 provided to the commission („PatientFSession- Details“) - 4 (of 6) „feedback sessions“ und 1 (of 3 based on the article, p. 7, Fig. 2 or 2 based on p. 6, Table 1) „open question sessions“. Even if the ex post constructed matching of the various data and information were accepted, results are thus presented in the article for which no NIRS data are present.

Furthermore, the data pool that is present shows that the statement in the article, p. 17 („Three to four sessions were performed each day, ...“) is not correct. Thus on 24.03.2014 patient F had 5 sessions, on 15.08.2014 patient B had 5 sessions, and on 21.06.2014 patient G had 8 sessions and on 16.12.2014 patient W had 5 sessions.

Some of these data were not included in the paper as stated above but erroneously transmitted in April.

c) NIRS-data: lack of matching

aa) The NIRS data that were transmitted contain overall no information which NIRS data relate to which steps (training, feedback, open questions) in the experiments that were described in the article. The data also do not contain information which questions were given in the respective session and which answers were identified. This alone precludes a reconstruction which data were used for which steps in the experiments that were described in the article. Accordingly the data that were transmitted do not permit to evaluate how the results of the article were arrived at. This means that the results that were described in the article are not backed by data in a manner that is reproducible.

This is especially true for the data on the so-called „open question-sessions“ (see their number in the article p. 6, Table 1 or p. 7 Fig. 2). In these sessions open questions wwere

asked – based on the description in the article and a yes-no answer was identified. There should be data (in test form) from which one can deduce which questions were asked and which data can be assigned to the identification process for a specific question. Only then would a data set or data sets have the quality to contain data of an “open question-session“.

The data that were transmitted contain no information of this type. Even if one granted (erroneously) that the first two steps in the experiments served the development and the testing of the model, and the respective questions and answers would be present, it would still be necessary to combine the respective questions and identified answers. In the data that were transmitted this no such identifiable data to „open question- sessions“ could be found. Thus results are presented without data.

This is scientific misconduct according to § 1 Abs. 2 Nr. 1 a) (invention of data).

We disagree with this evaluation of scientific misconduct. We did not invent any data. We surely agree that we could have provided more extensive documentation including text files. However, this was never requested by the reviewers or the journal. We already provided these data to the committee and would be happy to upload them as additional information to the paper. As noted before, in the BCI field such detailed data documentation has so far not been the rule and we did not intentionally leave out this very detailed documentation.
We would also have gladly provided more data on the open questions had we been asked, however, we never received any additional requests.

bb) The associations between the NIRS data with additional information about the tables that were transmitted in May 2019 do not change this (cf., o. 2. c). On the one hand the concerned persons admit in their letter to the DFG on 17.05.2019 (there under 2.), which was forwarded to the head of the commission per Email on May 18, 2019, explicitly that

- For all feedback sessions of patient G
- And the feedback sessions of patient F

in visit 3 (04.-07.8.2014) no data exist for the respective yes and no answers

This is not true. We stated that for these sessions only computer-generated summary files were created that could in no way be manipulated by the experimenter. In fact, the percentage of answers exceeding the threshold per patients is lower in the summary files than the individual files, thus biasing the data against our hypothesis of communication ability.

The respective ReadMe-Datei shows for the new data for patient G: „As mentioned in the letter individual "yes" and "no" were not saved for this patient because of the glitch in the software so we just have percentage value for this patient.“ The explanation in the letter about the determination of the percentage right answers in these sessions is not clear. At least the computation cannot be followed- Thus the article contains data – based on the statement of the concerned persons – for which no data exist.

This is not true. In the letter to the DFG of May 17, 2019, we clearly stated that „The developed BCI software also had the provision of calculating the percentage of

correct answers. Thus, for sessions where the system could not save the individual answers, we just saved the total percentage of correct responses and later used that percentage as the result of the particular session". So, the committee had the information on how these trials were computed.

For patient F we have to state that for three of the „open question sessions"

- No NIRS data are available (cf. o. 3. b) a. E.) and
- For the two others no data record is available

Furthermore all information for the training sessions is missing.

This is not true. The committee could have determined that there was a discrepancy between our data transmission from April 2 and the May download and could have asked us about this on May 22 or May 29. Moreover, the expert, who must have noted the discrepancy, much earlier could also have requested this information from us. Moreover, information on the training sessions is available in the paper.

Finally, the statements of May 2019 are ex post facto conducted reconstructions. At the time of the publication of the article (January 2017) this assignment did not exist and could not have been presented.
Accordingly Dr. Chaudhary has transmitted the data materials without an intelligible relationship to the experiments and results presented in the article.

We cannot follow this argument. We still have the time stamped files from 2017 that contain all these data. We later assembled an EXCEL file for the DFG and the commission to make it easier for the commissions to understand the data.

cc) Even if one were ready to accept the newly compiled associations of the data from May 2019, the conclusion would not change.

In patient B two „open question sessions" were completed. For the first session „V2D4b5_QuestionList.txt" und „V2D4b5_result.txt", „V2D4b5_QuestionList.txt" is introduced by the following text:

> „Normally open questions were stored with an 003_ number name, but during this visit to the patient we had problem with the open question presentation codes so we randomly rename the open questions as true and false (which means 003_name was renamed as 001_name and 002_name--please see the OQ family folder inside the list of sentence folder for the proof). Hence the label here as [sic] no real meaning, therefore please note that label 0 and 1 are 2 in reality."

We apologize if this has been confusing but again, questioning us could easily have resolved the problem. On this day the program for asking open questions which has a different code than the program that asks questions with known answers did not run. You have to take into account that the teams travel several hundred miles to work with these patients for some days. The problem could not be fixed on site. Rather than interrupting the costly experiment and go home, the team decided to use the program that is normally used for known questions and the question files for this were renamed as described above. This was a totally acceptable procedure, since

this is only a formal operation and did not affect the data that were collected. To prove this we also provided a folder with a time stamp that shows the renaming and the original folder.

At this stage, we would like to explain how the accuracy of a feedback session and an open question session is calculated. Please see the paragraphs below:

In every session (training and feedback) 20 known questions are being presented for which a question list is created, this question list has 10 true and 10 false sentences in random order. The answer to false sentences is 0 (no) and the name of the sentence begins with 001_(number) while the answer to a of a true sentence is 1 (yes) and the name of the sentence begins with 002_(number). For feedback sessions the result file, which is being created for each session, consists of the answer to each question predicted by the classifier either as 0 (i.e., the classifier predicted as no and BCI said "your answer was no") or 1 (i.e., the classifier predicted as yes and BCI said "your answer was yes"). To calculate the percentage accuracy of a session the presented answer, i.e, the answer of the question in the question list file is matched with the label of the predicted answer in the result file, i.e., if the label of the presented question is 1 and also the predicted label is 1 then the answer was predicted correctly, while if the label of the presented question 0 and the predicted label is 1 then the answer was predicted wrongly.

For the open question session, since we do not know the answer of the question, the question list created has just one answer (we used the number 2 as answer in our program) for all the questions and the name of the sentence begins with 003_(number). As usual the BCI also creates a result file with the predicted answer for each question as 0 (i.e., the classifier predicted as no and BCI said "your answer was no") or 1 (i.e., the classifier predicted as yes and BCI said "your answer was yes"). This answer is then matched with the answer estimated by the family member to calculate the overall accuracy of the session.

When we encountered a problem with running the open question module of the BCI at the patient's bed-side we renamed the open question as known question, i.e., 003_(number) as 002_(number) and 001_(number) in a random order where 10 out of 20 open questions were renamed as 002_(number) and another 10 as 001_(number), to present the open questions using the known question code of the BCI and bypass the dysfunctional code. For analysis purposes, the answer of each open question was matched with the estimated answer of the family members to calculate the overall accuracy of an open question session.

It ist o be stated that the data were used despite a problem in data acquisition and data were renamed later. The cited text probably should indicate that in this data set (in contrast tot he respective overview in the „feedback sessions“) the beginning of the data name with 001_ or 002_ does not indicate „true sentence“ or „false sentence “ and that the value contained in the following line is not 1 for true and 0 for false. (cf., patient F, data set V2D6b5_QuestionList.txt:
„Open questions were stored with an 003_ number name. Since we do not [sic] the answer of the open question in advance so we use trigger 2 as the marker.)

See above, there was not any data manipulation.

Furthermore, „V2D4b5_QuestionList.txt“ contains the data names of die 20 Soun data files, „V2D4b5_result.txt“ has, however, *21* 0- or 1-entrues, respectively. For none of the questions a clear assignment is therefore possible since added value could be at any place in the number list.

The 0 and 1 indicated in the result file for the open questions pertain to the answers generated by the classifier using the software for the known questions and an extra answer was added by the software because of an error. However, as this was an open question session, the 0 and 1 answers assigned to the feedback sessions had no meaning and were not analyzed since open question sessions do not have known yes/no answers. We described this in the beginning of the result text file that was sent to the DFG on May 17 and downloaded by the commission on May 20. As noted there the contents of the file has no real meaning because the open question answer was matched with the estimated response of the family member in real time and these data were entered and used.

„V2D4b5_result.txt“ is introduced with the following text: „The accuracy of open question session is an estimation as written in the manuscript and is Based [sic] on the feedback of the family members.“ What follows are 14 1- and 7 0-entries. This would yield a „classifi-cation accuracy“ of 14/21, which is 66.67 % (if one assumes 20 cases, 70%). In the article, p.9, fig. 4 much higher values are presented.

The classification accuracy of an open question session is being calculated by matching the predicted label of each answer predicted by the classifier (either as 0 i.e., the classifier predicted as no and BCI said “your answer was no” or 1 i.e., the classifier predicted as yes and BCI said “your answer was yes”) and the answer estimated by the family member. For example, if the classifier predicts 0, which means “No”, and the family member also estimates that the answer is “No” then that was taken as correctly classified answer and so on. The label matching which the committee members are talking about is very well valid for a known question feedback session but the committee member has neglected the fact it was an open question session run as a feedback session which means that they are matching the meaningless label of a question, i.e., 0 or 1 (which in reality is 2) with the predicted label, 0 or 1. As written in the answer above, at this point we would like to reiterate that:

For the open question session, since we do not know the answer of the question, the question list created has just one label (we used number 2 as label in our program) for all the questions and the name of the sentence begins with 003_(number). As usual the BCI also creates a result file with the predicted answer for each question as 0 (i.e., the classifier predicted as no and BCI said "your answer was no") or 1 (i.e., the classifier predicted as yes and BCI said "your answer was yes"). This answer is then matched with the answer estimated by the family member to calculate the overall accuracy of the session.

For the second „open question session“ in patient B the overview table gives „session“ V2D5b5 . Present are, however, tables „V2D5b6_Question- List.txt“ and „V2D5b6_result.txt“. Thus there is no information that permits an association if the matching data.

When renaming the file name (to make sure that the reviewers at the DFG would be able to associate the files to the respective case), there was an inadvertent change of the number: we apologize for the oversight. The file name should have been …..05.

„V2D5b6_result.txt“ contains (after the same introductory sentence) moreover 16 0- und 4 1- entries. This would yield a „classification accuracy“ of 4/20, i.e. 20 %. In the article, p. 9, fig. 4, much higher values are presented.

As explained above this pertains to a an open question session which was run as a known question feedback session after renaming the open question files as explained above wherein the 003_(number) (the file name used for an open question) was renamed as 002_(number) and 001_(number) (the file name used for a known question) in a random order to bypass the dysfunctional open question module of the BCI software. Also as explained above “The classification accuracy of an open question session is being calculated by matching the predicted label of each answer predicted by the classifier (either as 0 i.e., the classifier predicted as no and BCI said “your answer was no” or 1 i.e., the classifier predicted as yes and BCI said “your answer was yes”) and the answer estimated by the family member. For example, if the classifier predicts 0, which means “No”, and the family member also estimates that the answer is “No” then that was taken as correctly classified answer.” The percentage value thus reported was based on the matching of these labels and we reaffirm that it is correct.

For the respective tables for the two „open question sessions“ for patient B the information about the answers that were given by the BCI system, which can then be associated with the estimated accuracy of the family members, was missing.

As explained above the accuracy of an open question session was calculated as follows: “The classification accuracy of an open question session is being calculated by matching the predicted label of each answer predicted by the classifier (either as 0 i.e., the classifier predicted as no and BCI said “your answer was no” or 1 i.e., the classifier predicted as yes and BCI said “your answer was yes”) and the answer estimated by the family member. For example, if the classifier predicts 0, which means “No”, and the family member also estimates that the answer is “No” then that was taken as correctly classified answer and so on.” As mentioned in page 13 of our manuscript Chaudhary et al., 2017 “Still, we have to remain cautious about our judgements to open questions' answers”
And again on page 18 of our manuscript Chaudhary et al., 2017
“ The validity of answers to open questions can only be estimated by (a) face validity (i.e., questions of pain in the presence of an open wound); (b) stability over time; (c) external validity, estimated by family members and caretakers; and (d) internal validity between questions (i.e., the concordance between the answer to “I love to live” with the answer to “I rarely feel sad” [presented to all patients -except W-regularly]).”

Thus, it can be seen that calculation and meaning of the open question classification accuracy was explained in our original publication which the University commission did not use.

Overall, in the presented case, even if the ex post associations that was given by the concerned persons were accepted, the article presents results without data.

This is not true. We did not create any files. We just did not provide as detailed information as the commission members request now but we can provide this information if requested. The reviewers of PLoS Biology did not request it.

Further inconsistencies are present in that the two QuestionList-Data sets point to sound-data from the same day although the respective sessions were performed on two different days. In addition, the time stamp of the listed sound data is in opposition to the listed sequence in which they were used.

All the sound files were created at different time points before the sessions. This could also have been clarified in a discussion with the commission.

4. Potential data falsification by errors in data analysis

a) In the transmitted data two scripts are included, „NIRs_trainmodel1.m“ and „trainSVMlinearclassifier.m“ (for the latter see Report, attachment 1A, Sp.10, text of. 9.10.2017, Appendix A.2). „NIRs_trainmodel1.m“ loads the NIRS-data und does preprocessing of the raw data. At least if the variable „feature“ hast he values 2 or 3, in lines 204 and 221 the function „fea_csptrain“ is called up, which probably extracts the relevant features from the raw data. This for the work of the commission (potentially) extremely important function was, however, not made available. In line 227 „NIRs_trainmodel1.m“ then calls up „trainSVMlinearclassifier.m”. The results that are obtained by the use of this script are saved by „NIRs_trainmodel1.m“ (lines 229-245).

The script „trainSVMlinearclassifier.m“ conducts a parameter optimization with cross validation („cross-validation“) (lines 12 and 15, flag „-v fold“). There “classification accuracy“ includes a cross-validated comparison of numerous parameter combinations and the combination with the highest „classification accuracy“ is determined. For the best combination of hyperparameters, „i“ und „k“ are then again trained in a linear „Support Vector Machine“, but without cross-validation (line 23). Without use of the statistical method of cross-validation, a model that is obtained by optimization and applied to the same data set on which it was optimized, leads to a distortion of the results in the direction of significant results. Only the application to an independent set of data will prevent distortion from occurring (this is the essence of cross-validation).

This cannot conclusively be tested by the current script. Would the model that has been obtained in line 23 be applied to new online-data, the procedure would be correct. Are the values that are reported in the article those that were obtained by the optimization in line 23, then they are with a high likelihood to high („bias“). The MATLAB-scripts needed for the evaluation of this question were not provided to the commission.

The commission has correctly understood that the model was built on cross validation and that the subsequently generated optimization with the support vector machine was in fact applied to new data (feedback sessions). We would like to point out that uploading this type of script has not been the norm so far in BCI articles (for example “Hochberg, Leigh R., et al. "Neuronal ensemble control of prosthetic devices by a human with tetraplegia." Nature, 442.7099 (2006): 164.“; Hochberg, Leigh R., et al. "Reach and grasp by people with tetraplegia using a neurally controlled robotic arm." Nature485.7398 (2012): 372., among many others) due to intellectual property issues and was also not requested by PLoS Biology. We would have been happy to provide the script had we been asked to do so by the commission.

b) Independent of this, the report of the ombudspersons suggests the following: the concerned persons should have known about the problem („bias) at the latest with the detailed „report“(Report, attachment 1A) transmitted per email on 06/10/2017. The communication that followed in short time afterwards shows that the concerned persons admit it themselves, Email of. 16.10.2017 (16:18) of Prof. Birbaumer (Report, attachment 1G, p. 1): „…: we used a hypothesis driven feature and model building, which is statistically not correct but physiologigcally [sic] plausible.“ Along these lines it was pointed out in the email of Prof. Birbaumer of 16.10.2017, (16:54) (Report, attachment 1H, p. 1): „the model we built was based on our a priori hypothesis how it should look like ...“ In the email of Prof. Bir- baumer of 16.11.2017, (Report, attachment 1L, S. 1) it is written::

> „Yes, we work on the exact correction of mistakes, I want to wait fort he independent video analyses … you are absolutely right, mistakes must be corrected, and we will do it. But I do not want to publish a paper or a commentary that is too negative, for clinical reasons we must show an alternative that works. I do not know if a published paper can be retracted…….if you withdraw your commentary, we could make a correction paper together, but as long as the commentary is sent without our discussion and common understanding, then not.“

In the course of the procedure the commission has, moreover, been informed that Prof. Birbaumer knew already on 5.11.2015, that is more than one year before the publication via email by a former co-worker that the data did not yield significant results if they were computed in the right manner but that a correct statistical analysis would yield a statistical normal distribution. This information was credible and substantiated by figures from documents.

Prof. Birbaumer made these statements within a number of discussions about the best way to analyze these patient data. At this time, as noted above, there were many discussions about the role of physiologically plausible models and purely data driven models. Prof. Birbaumer, during his whole career, intended to model a Socratic attitude to his students: attend positively to all criticisms and doubts at all time during the development of a report. Thus, he took the criticisms at that point in time seriously and then discussed it with all the collaborators and critics. He was not aware at that time of a mail message and of all the data and calculations done. What Prof.

Birbaumer intended to focus on where physiologically plausible models as stated in the a priori hypotheses. He also emphasized, particularly to scientists and students from other disciplines like engineering, informatics and psychology that any question, any critic and interpretation needs to respect and understand the physiological basis of the acquired data, in this case NIRS and EEG. Statements of this kind cannot be taken as proof of acceptance of any type of specific results but served to stimulate discussion and re-reviewing on the part of his collaborators.

We are not aware and cannot document acceptance of any report of this sort in 2015. The only persons who analyzed these data in detail at the time were Ujwal Chaudhary and Bin Xia. We never gave these data to other persons for analysis until later in 2017. We are therefore unable to further comment on this and must refute this claim as unsubstantiated.

Therefor from this timepoint on in 2015 one can no longer speak of an error in the sense that the relevant methods were not mastered or the lack of statistical knowledge (see o. at Fn. 1).

These informations suggest that a falsification of data took place.

We have to refute this statement because we were not aware of any problems in 2015 and we only wanted to be careful in refuting or accepting alternate hypotheses about the data prematurely in the face of the very significant clinical and ethical issues involved in later discussions.

c) In addition, the commission notes the following: in the article, p. 19 it is written: „If the classifi- cation accuracies … were greater than the chance-level-threshold, a new model was ge- nerated using the relative change in O_2Hb across three training sessions to give online feedback.“

In the excel-overview for patient B („PatientB_SessionDetails“) the following is written about „Feedback-session“ „V2D4b3“ (15.08.2014): „The model was built using the V2D4b1 and V2D4b2.“ Thus the model was built on the basis of only two „training sessions“.. The same is true for the „Feedback-session“ „V2D5b3“ (16.08.2019), here also the model was based on only two „training sessions“ („V2D5b1“ und „V2D5b2“).

The same is true for patient F for the „feedback-sessions“ „V2D6b3“ (20.05.2014) and „V3D3b3“ (06.08.2014), whereas „V2D5b4“ and „V3D4b4“ actually contain three „training sessions“. For patient G only once the model was based on three „training sessions“, otherwise on two, for patient W only twice based on two training sessions (besides, once based on three and once based on five). Thus the model was based on two, three or five “training sessions“. This is in direct opposition to the statement in the article cited above, p. 19.

In some cases, only 2 rather than 3 training sessions were used. This was the case when the model yielded good differentiation already after two sessions, which was advantageous for the patients who cannot train for extended periods of time. Sometimes more training sessions were needed. We used the best available classification model, which sometimes emerged early, sometimes later but was on average 3 sessions. We meant to write an average of three training sessions and we can correct this in the paper to average of 3 training sessions, if requested.

5. Participation in the review process of the „formal comment“

Dr. Spüler doubted the effectiveness of the methods the authors used in their articleand formulated this in a „formal comment“ that was published in PLOS-Biology on 08.04.2019 (https://doi.org/10.1371/journal.pbio.2004750). This text was first returned by PLOS Biology for revision in December 2017 based on the statements of two reviewers, the revised text was then rejected in March 2018 (Cf. Report, attachment 2B) and published only after a rebuttal of Dr. Spüler against this decision (cf. Report, attachments 3 and 4).

After the opening of the investigations, Prof. Birbaumer sent several data sets to the commission, among them a 11-page text, which is called „Review“ in the first sentence („This is a review of the Comment … by Martin Spüler ...“). In the above mentioned review process Prof. Birbaumer also called his statements „review“ and „re-review“. This could give the impression that he was involved in the review process despite a conflict of interest , what could be relevant with respect to § 1 Abs. 2 Nr. 2 g) rules of procedure (intentional delay of the publication of a scientific work by a reviewer).

However, this is a wrong nomenclature. Prof. Birbaumer did not take part as an (anonymous) reviewer in the review process, but has given a statement where he stated his name (cf. Report, attachment 2B, email of the editor of 13.12.2017: „Your manuscript has been evaluated … by an Academic Editor with relevant expertise, and by two independent reviewers … In addition, Dr Niels Birbaumer provided signed comments ...“; according to the email of 12.03.2018, Report, Attachment 3, p. 2).
Therefore he was not acting as a „reviewer“ in the sense of the rules of procedure but as author of the publication, on which the criticism of the “Formal Comment“ was based.
A scientific misconduct according to § 1Abs. 2 Nr. 2 g) of the rules of procdure was thus not realized.

Both the whistleblower and Prof. Birbaumer reviewed each other‘ s commentary which was required by PLoS Biology and the term review refers here to a thorough reading and evaluation of the respective paper not a formal review process. The word review has both meanings in the English language.

A) Recommendations of the commission

Based on § 16 S. 2 the commission can make recommendations for the further procedure, whereby type and severity of the misconduct and rights and interests of third parties must be considered.

I. In the present case the scientific misconduct has not only science immanent consequences. Rather, the affected patients, their caretaking family members, and at least one health insurance company are affected, which was sentences to finance the equipment that is needed for the method the concerned persons developed. By the fact that the procedure became public, the reputation of the scientific research the public domain has also been damaged. Societal trust in scientific research has been disappointed. This damage has also to be accounted for by the concerned persons who knew about the incorrectness of their method (cf. o. A) III. 4. c)).

II. Under inclusion of these ts the commission makes the following recommendations:

1. The concerned persons Prof. Birbaumer and Dr. Chaudhary have to be imposed to withdraw the publication.

2. The editors of the journal „PLOS Biology“ have to be informed about the verdict and have to be asked independently of this to withdraw the article since it is a serious breach against their own “Data Availabilty Policy”.

3. The involved funding agencies (such as DFG, Volkswagen-Stiftung, Bundesministerium für Bildung und Forschung, Baden-Württemberg-Stiftung, Eva Luise und Horst Köhler-Stiftung, cf. Article p. 1f., li. Sp.) have to be informed about the verdict.

4. The central association of the statutory health insurances (GKV-Spitzenverband) as well as the central association of the private health insurances (Verband der Privaten Krankenversicherung e.V.) have to be informed about the verdict.

5. The rectorate should determine if the statements made in the verdict have consequences for the – in any case time limited – award (see directive of the rectorate of 9.11.2011, Nr. 1) of a senior professorship to Prof. Birbaumer. Independent of this the commission recommends to the rectorate, to avoid to further use the procedure named in Nr 4 of the directive for Prof. Birbaumer.

6. The inconsistency of the data materials given to the commission requires further to have an external review of all publications in which the concerned persons were involved since the acquisition of these data in 2014 .

7. The scientific cooperation partners of Prof. Birbaumer should be informed about this verdict in an appropriate manner to give them the opportunity to review their research work.

8. In accordance with the “responsibility” that has been enshrined in the mission statement of the university the commission suggests to offer a point of contact to the relatives of the patients, where appropriate.

III. Finally the commission wants to point out the specific aspects of ethics in research that

are associated with this case and wants to encourage the university to initiate a respective discussion in the concerned research institutions – outside of those institutions involved with potential cases of scientific misconduct .

The research group of Prof. Birbaumer claims from a lay person's point of view that the system the group developed can interpret brain data of patients in a Completely-Locked-in-State in a self-learning fashion and can translate this in yes-no statements. In this way information from the patients can be assessed in a reliable, because scientifically based, manner and be exchanged with them. The research group thus presents a scientifically valid way to communicate with patients in a Completely-Locked-in-State. The evidence for the affirmed communication is solely the scientific validity of the procedure. The yes/no statements of the patients cannot be determined in any other way than with the BCI and the contents cannot be tested in any other way, with the criterion being the scientific quality of the used system. Thus the research group has the responsibility to guarantee the scientific evidence of its research results and to document its research results with all necessary information about the research method and the research process.

This responsibility exists not only towards the scientific public but also in a special way towards the patients who suffer from ALS who approach a Completely-Locked-in-State and who have to rely on the promise that they can communicate in this state with the help of a schientifically qualified system as well as towards the people who are caregivers to the patients in this state. Patients and those who are close to them cannot themselves test the promise that they can communicate with each other. They have to solely rely on the scientific basis of the promise.

With the „size" of the medial dissemination of the promise the research group also has increased the responsibility for its scientific proof and also has a responsiblity that it creates the precondituons that ist published results can be tested and thus the affirmation of communication in patiens with Completely-Locked-in-State can be verified or falsified.

Tübingen, 30.05.2019

Xxx (head of commission)