In diesem Dokument kommentieren wir umfänglich die Entscheidung der Kommission zur Untersuchung von wissenschaftlichem Fehlverhalten der Universität Tübingen. Wir glauben, dass die Kommission ihr Urteil auf Daten basiert hat, die falsch übertragen wurden und dass die Kommission zusätzliche Daten, die wir ihr übergeben haben, nicht berücksichtigt hat, sowie andere Aspekte unserer Forschung missverstanden hat. Ein Austausch über diese Vorgänge wurde von der Kommission nicht gesucht. Wir möchten auch darauf hinweisen, dass die Kommission nicht berücksichtigt hat, dass unser Artikel und der Kommentar und die hochgeladenen Daten in PLoS Biology von insgesamt 8 Gutachtern wie auch von den Herausgebern von PLoS Biology begutachtet worden sind. Darüber hinaus wurde der Umstand, dass unsere Daten von einem unabhängigen Experten für maschinelles Lernen, Dr. Sudhir Pathak, repliziert und publiziert wurden, nicht beachtet.

Universität Tübingen
Kommission zur Untersuchung von Fehlverhalten in der Wissenschaft

Aufgrund des Berichts der Vertrauenspersonen der Medizinischen Fakultät (xxx, xxx) vom 22.11.2018 hat die Kommission in der Sitzung vom 23.1.2019 hinsichtlich der Betroffenen Prof. Birbaumer und Dr. Chaudhary ein Verfahren eröffnet.

Am Verfahren beteiligt (§ 11 VerfahrensO) haben sich die Vertrauenspersonen xxx, xxx (bis zum Zeitpunkt der Niederlegung ihres Amtes im Februar 2019), xxx, xxx und xxx.

xxx schied am 01.04.2019 aufgrund eines Forschungsfreisemesters aus der Kommission aus. Er wurde anschließend durch xxx vertreten. In der Sitzung
v. 17.04.2019 wurde xxx durch die Kommission als Sachverständiger bestellt (§ 13 VerfahrensO).

Wir beraten uns derzeit mit einem Anwalt, inwieweit die Zusammensetzung der Kommission und die Wahl des Experten korrekt waren und in welchem Umfang Regeln der guten wissenschaftlichen Praxis befolgt wurden und kommentieren nicht weiter zu diesem Abschnitt. Wir möchten jedoch darauf hinweisen, dass der Experte xxx zunächst Mitglied der Kommission war, dann als externer Sachverständiger bestellt wurde und dass er demselben Institut wie der Hinweisgeber entstammt und der Anschein der Befangenheit nach den Regeln der guten wissenschaftlichen Praxis nicht auszuschließen ist.

Am 25.02.2019 wurde der Betroffene Prof. Birbaumer, am 06.03.2019 wurde der Betroffene Dr. Chaudhary entsprechend § 12 S. 1 1. Halbsatz der Verfahrensordnung angehört. Der Hinweisgeber wurde in einer Sitzung am 27.02.2019 befragt, der Vorsitzende der Kommission führte mit ihm außerdem am 02.04.2019 ein Telefongespräch.

Auf die Anhörung nach § 12 S. 1 2. Halbsatz der Verfahrensordnung haben beide Betroffenen verzichtet (jeweils mit Email v. 23.05.2019).

Es ist zutreffend, dass wir die Gelegenheit für eine Anhörung vor der Kommission nicht wahrgenommen haben. Wir hatten zu diesem Zeitpunkt keine anwaltliche Beratung und befanden uns im Irrtum über die rechtliche Situation. Wir missverstanden den Hinweis von xxx, dass die Anhörung ein

notwendiger Schritt sei, wie er von der Verfahrensordnung vorgesehen sei und dass die Kommission keine Fragen hatte, möglicherweise irrtümlich auf eine Art und Weise, dass es kein Interesse seitens der Kommission gab, unsere Stellungnahme zu hören. xxx schrieb, dass die Kommission die Daten, die wir am 17. Mai 2019 an die DFG übergeben hatten, heruntergeladen und diskutiert hatte. Wir können bestätigen, dass die Daten am 20. Mai 2019 heruntergeladen wurden. Da diese Excel-Dateien umfängliche Information über die Daten enthielten, sind wir unsicher, ob die Kommission ihren Inhalt tatsächlich umfassend gewürdigt hat, ohne dass sie diese Daten an einen Experten übergab und haben vermutet, dass die Kommission nicht wirklich an einer umfänglichen Analyse des Problems interessiert war. In Anbetracht der Komplexität des Vorgangs waren wir darüber hinaus nicht überzeugt, dass die Kommission unsere Antworten und Argumente umfassend gewürdigt hat. Wir hatten den Eindruck, dass die Kommission unter großem Zeitdruck handelte, da eine ausreichende Evaluation durch einen externen Sachverständigen in so einer komplexen Sache Monate dauern sollte, die Expertise jedoch tatsächlich in 4 Wochen erstellt wurde. So dauerte die Begutachtung des Artikels z.B. ein Jahr und unsere Datenerhebung und –analyse und das Schreiben des Artikels umfassten einen Zeitraum von insgesamt 3 Jahren. Dieses Vorgehen verstärkte unseren Eindruck, dass die Kommission sich nicht wirklich im Detail mit dem sehr komplexen Sachverhalt befasst hatte. Wir sahen deshalb keinen Sinn darin, mit der Kommission noch einmal zu sprechen, insbesondere, da xxx schrieb, „um Missverständnisse auszuschließen: Die Kommission hat keine weiteren Fragen oder ähnliches. Die Anhörung ist optional.“ Nachdem wir uns rechtlich beraten haben, verstehen wir jetzt, dass wir hier einen Fehler gemacht haben.

In der Sitzung vom 29.05.2019 fasste die Kommission einstimmig den folgenden Beschluss.

A) Beschluss

Ein wissenschaftliches Fehlverhalten der Autoren Prof. Birbaumer und Dr. Chaudhary liegt vor.

I. Sachverhalt

Die Betroffenen haben als gemeinsame Autoren den Artikel „Brain-Computer Interface-Based Communication in the Completely Locked-In State“ in der Online-Zeitschrift PLOS Biology“ (ISSN 1544-9173; eISSN 1545-7885) veröffentlicht (DOI:10.1371/journal.pbio.1002593; Tag der Veröffentlichung: 31.01.2017).

Diese Veröffentlichung beruht auf Daten, die während mehrerer Sitzungen mit vier Patienten erhoben wurden, welche in fortgeschrittenem Stadium an Amyotropher Lateralsklerose (ALS) erkrankt sind. Diese Patienten sind aufgrund des totalen Verlustes der Bewegungssteuerung, selbst der Augen und der Augenlider, nicht mehr in der Lage, mit ihrer Umgebung zu kommunizieren. Ihr Zustand wird daher als „Completely Locked-In State“ (CLIS) bezeichnet.

Dabei kam die Methode der funktionellen Nahinfrarotspektroskopie (fNIRS) zur Anwen-

dung. Unter Anwendung von Techniken maschinellen Lernens wurde versucht, ein Modell zu entwickeln, das erlaubt, bei der Anwendung auf einen Datensatz eine statistisch valide Zuordnung der jeweiligen Muster der Hirnaktivität eines Patienten zu einer (vom Patienten gedachten) „Ja“- oder „Nein“-Antwort vorzunehmen. In diesem Sinn ist von einem „Brain-Computer-Interface“ (BCI) die Rede.

In der Veröffentlichung wird behauptet, dass über diesen Aufbau mit einer deutlich über der Zufallswahrscheinlichkeit liegenden Wahrscheinlichkeit die Hirnaktivität eines Patienten einer (vom Patienten gedachten) „Ja“- oder „Nein“-Antwort zugeordnet werden kann (Artikel, S. 1: „Online fNIRS classification of personal questions with known answers and open questions ... resulted in an above-chance-level correct response rate over 70 %.“). Auf den Bericht, S. 2-3, wird verwiesen.

Die Arbeitsgruppe, die von Prof. Birbaumer geleitet wurde, hat Daten dem späteren Hinweisgeber Dr. Spüler zur Verfügung gestellt. Der Hinweisgeber ist selbst wissenschaftlich auf dem Gebiet des BCI tätig und kooperierte seit mehreren Jahren mit der Arbeitsgruppe um Prof. Birbaumer in diesem Bereich. Auf den Bericht, S. 4, wird verwiesen.

Dr. Spüler hat Zweifel an der Leistungsfähigkeit der von den Autoren in ihrem Artikel vorgestellten Methode in einem „Formal Comment“ formuliert. Dieser Text wurde von PLOS Biology zunächst auf Grund der Stellungnahme von zwei Gutachtern im Review-Prozess im Dezember 2017 zur Überarbeitung zurückverwiesen, der überarbeitete Text wurde dann im März 2018 abgelehnt (vgl. Bericht, Anlage 2B). Auf Widerspruch von Dr. Spüler gegen diese Entscheidung wurde sein überarbeiteter Text am 08.04.2019 publiziert (https://doi.org/10.1371/journal.pbio.2004750) (vgl. Bericht, Anlagen 3 und 4). Gleichzeitig wurde die „Response to: ‚Questioning the evidence for BCI-based communication in the complete locked-in state‘“ von Prof. Birbaumer und Dr. Chaudhary publiziert (https://doi.org/10.1371/journal.pbio.3000063).

Im Anschluss an seine Anhörung am 06.03.2019 hat die Kommission von Dr. Chaudhary die Herausgabe aller Daten, die für den Nachvollzug der im Artikel gemachten Schritte, so wie er sie in seiner Präsentation dargestellt hatte, nötig sind, verlangt (Email v. 28.03.2019). Dr. Chaudhary hat daraufhin Daten im Umfang von ca. 5,2 GB zur Verfügung gestellt. Es handelt sich um die NIRS-Daten zu den vier Patienten F, G, B und W, die jeweils in einzelne Ordner „visit [n]“ und mit einem Datum bezeichnete Unterordner (z.B. „2014-06-10“) eingeordnet sind, sowie zwei zur Auswertung benutzte Skripte.

Prof. Birbaumer hat dem Vorsitzenden der Kommission mit Email v. 18.05.2019, in welcher er das an die DFG gerichtete Schreiben der Betroffenen v. 17.05.2019 mitgeteilt hat, einen Link gesandt, über welchen (allerdings erst nach vorheriger Autorisierung durch die Betroffenen) weitere Dateien zum Download zur Verfügung gestellt wurden („Additional Data for DFG“, enthaltend das Archiv „17052019_Reply.zip“). In dem Schreiben sowie den zur Verfügung gestellten Dateien werden das Fehlen von Daten, der Ausschluss von Sitzungen von der Auswertung sowie die spätere Umbenennung von Dateien, die während der Experimente erstellt worden waren, eingeräumt. Außerdem wird beschrieben, dass es sich bei einigen „training sessions“ um „Text validation sessions“ handelte, die im Artikel (S. 17) nicht beschrieben werden.

Wir verwahren uns entschieden gegen diese Interpretation unseres Schreibens an die DFG. Wir hatten den Eindruck, dass einige der Fragen, die der Hinweisgeber

aufwarf, aus einem Missverständnis der Beschreibung der Daten und Prozeduren resultierten, welches von uns nicht intendiert war. Wir bemühten uns deshalb, die Prozedur und was wir genau getan haben, in einer sehr viel detaillierteren Weise zu beschreiben, als es von der Zeitschrift verlangt wurde. Diese zusätzliche Information sollte jetzt nicht gegen uns ausgelegt werden, insbesondere, da sie nichts an dem Algorithmus, den wir verwendet haben, um die Kommunikationsfähigkeit der Patienten zu bestimmen, ändert. Sie dokumentiert nur die Probleme, die in der Interaktion mit dem Patienten auftraten in einem sehr detaillierten Umfang und legt dar, was in jeder Sitzung im Detail passierte. Die zusätzliche Information verdeutlicht auch detailliert den Einschluss und Ausschluss der Trainingssitzungen, die wir verwendet haben, um das Klassifikationssystem zu etablieren. Sie dokumentiert auch die Feedback-Sitzungen, die verwendet wurden, um die Kommunikationsfähigkeit zu bestimmen. Während die Patienten in den Trainingssitzungen, d.h. in den Sitzungen, in denen das Klassifikationssystem gebildet wurde, keine Rückmeldung darüber erhielten, ob die Antwort richtig oder falsch war, erhielten die Patienten in den Feedbacksitzungen, die zur Bestimmung der Kommunikationsfähigkeit der Patienten verwendet wurden, nach jeder Frage Feedback über die Richtigkeit ihrer Antwort.

Wir verwendeten in der Analyse alle Durchgänge, um die Kommunikationsfähigkeit zu bestimmen, sogar wenn die Patienten ja und nein nicht differenzieren konnten. Dies war in 7 von 21 Sitzungen der Fall, in denen wir die Kommunikationsfähigkeit analysierten und der Prozentsatz der korrekten Antworten basierte auch auf diesen Sitzungen. Die Sitzungen, die von uns Textvalidierungs-Sitzungen genannt wurden, waren normale Trainingssitzungen und wurden wie alle anderen Trainingssitzungen behandelt. Es gab deshalb keine Notwendigkeit, diese Sitzungen separat in dem PLoS Biology Artikel von 2017 zu erwähnen. Wir erwähnten diese Sitzungen jetzt nur, weil wir glauben, dass es möglich ist, dass der Hinweisgeber diese Bezeichnung der Sitzungen in den Daten, die wir ihm übergeben haben, gelesen hat und diese fälschlicherweise als Feedback-Sitzungen identifizierte, die wir verwendet haben, um den Prozentsatz korrekter Antworten für die Kommunikationsfähigkeit zu ermitteln. Jedoch hatten die Textvalidierungssitzungen nur die Eigenheit, dass die Experimentatoren das Ergebnis der Prädiktion der Antwort sehen konnten und somit schauen konnten, ob die vom Gehirn vorhergesagte Antwort und die bekannte Antwort übereinstimmten. Es waren keine Feedback-Sitzungen und sie wurden nicht verwendet, um die ja-nein-Prozentsätze zu berechnen und damit die mögliche Kommunikationsfähigkeit der Patienten zu ermitteln. Wie wir bereits festgestellt haben, hat eine unabhängige Berechnung der Daten von Dr. Sudhir Pathak sogar höhere Einschätzungen der Kommunikationsfähigkeit ergeben.

Wir möchten feststellen, dass die Daten, die wegen einer Fehlfunktion der Gerätschaft ausgeschlossen wurden, in keinem Artikel, den wir kennen, je berichtet werden und sie hätten auch nicht analysiert werden können, da in diesen Fällen die Trigger nicht richtig zugeordnet werden konnten. Daten wurden niemals von uns umbenannt, sondern nur der Name des Satzes in einer Textdatei, die für das Programm verwendet wurde. Sätzen mit offenen Fragen wird eine 003 vorangestellt und die Antwort in dem Satz wird jeweils durch eine 2 dargestellt, da es unbekannt ist, was die richtige Antwort sein könnte. Im Gegensatz dazu wurden Textdateien für Sätze mit bekannten Antworten eine 001 oder 002 vorangestellt, je nachdem, ob der Satz eine richtige oder eine falsche Aussage enthielt und die korrekten Antworten auf die Sätze wurden dann mit 1 (ja) und 0 (nein) benannt. In dieser spezifischen

Sitzung wurden offene Fragen benutzt, aber in einer Datei für Feedback-Fragen angewandt. Dann wurde das normale Vorgehen für die Berechnung der prozentual richtigen Antworten für offene Fragen verwendet. Es wurden somit keine Dateninhalte verändert. Wir glauben, dass dies eine akzeptable Prozedur ist, da es sich nur um eine formale Operation handelte. Um dies zu belegen, haben wir auch die Datei mit einem Zeitvermerk übergeben, der zeigt, dass die Umbenennung zum damaligen Zeitpunkt passierte, sowie die Originaldatei. Es waren jeweils 2 bis 3 Personen beim Patienten anwesend, die dieses Vorgehen bestätigen können. Darüber hinaus wurden Sitzungen mit offenen Fragen nie verwendet, um die Kommunikationsfähigkeit der Patienten zu dokumentieren – sie wurden nur verwendet, um zu zeigen, wie sie angewendet wurden und um die Leser zu ermutigen, sich zu überlegen, solche Fragen auch bei den Patienten zu verwenden, um deren Wünsche zu ermitteln. Somit hat dies keine Konsequenz für die Feedback-Sitzungen, die wir verwendet haben, um die Kommunikationsfähigkeit der Patienten zu ermitteln.

II.
III.Zum Prüfungsmaßstab

Der Begriff des Fehlverhaltens in der Wissenschaft wird in § 1 VerfahrensO beschrieben. Dabei werden in § 1 Abs. 2 zu drei Fallgruppen (Falschangaben, Verletzung geistigen Eigentums, Beeinträchtigung der Forschungstätigkeit anderer) im Einzelnen bestimmte Verhaltensweisen benannt.

Im Gegensatz zur Verfahrensordnung zum Umgang mit wissenschaftlichem Fehlverhalten der DFG beschränkt sich die Verfahrensordnung der Universität Tübingen jedoch nicht auf diese drei Fallgruppen, sondern definiert in § 1 Abs. 1 Fehlverhalten als „Verhalten in einem wissenschaftserheblichen Zusammenhang, das gegen Rechtsvorschriften, oder gegen solche geschriebenen oder ungeschriebenen Regeln verstößt, deren Einhaltung allgemein, in einem bestimmten wissenschaftlichen Fach oder einer wissenschaftlichen Einrichtung als unabdingbar angesehen wird.“ Die Fallgruppen des § 1 Abs. 2 stellen also, wie sich auch aus dem Wortlaut „insbesondere“ ergibt, keine abschließende Aufzählung dar.

Entsprechend hat die Kommission ihre Untersuchung nicht auf die Fallgruppen in § 1 Abs. 2 beschränkt, sondern auch untersucht, ob für den vorliegenden Fall Regeln verletzt wurden, deren Einhaltung als unabdingbar angesehen wird.

Andererseits ergibt sich aus der Regelung in § 1 Abs. 1 und 2 der Verfahrensordnung, dass nicht jeder wissenschaftliche Fehler ein Fehlverhalten ist. Wissenschaftliches Fehlverhalten ist ein Fall von Unredlichkeit, nicht von Irrtum.[1] Irrtümlich methodisch verfehlt angelegte Experimente, Denkfehler, irrtümlich falsche oder unterlassene Anwendung einschlägiger statistischer Methoden usw. sowie unklare, unschlüssige oder in sich widersprüchliche Aussagen stellen also nicht als solche wissenschaftliches Fehlverhalten dar. Entsprechend obliegt die Qualitätssicherung in einem bestimmten Forschungsbereich der jeweiligen wissenschaftlichen Gemeinschaft; die Kommission Fehlverhalten ist hierzu nicht berufen und darf sich für diesen Zweck auch nicht zur Verfügung stellen. Die Frage, ob und inwieweit solche Mängel vorliegen, kann daher grundsätzlich nicht Gegenstand einer Untersuchung durch die Kommission Fehlverhalten sein.

Daher war von den im Bericht, S. 5, in einer in fünf Absätze aufgeteilten Liste dargestellten Sachverhaltskomplexen der im vierten Absatz angesprochene (in sich nicht nachvollzieh-

bare Ausführungen in der Veröffentlichung) von vornherein nicht Gegenstand des Verfahrens.

IV. Vorliegen eines Fehlverhaltens

1. Selektive Datenauswahl bei der Datenerhebung

a) Im Artikel heißt es auf S. 17 (d.h. in der Beschreibung des Experiments):

> „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 terminated 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 sessions, ...“

1 Vgl. Denkschrift der DFG „Vorschläge zur Sicherung guter wissenschaftlicher Praxis“ (1998, ergänzt 2013), S. 13 und 40.

Im Artikel heißt es auf S. 9: „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 der Email v. 9.10.2017 schreibt Prof. Birbaumer dagegen: „certainly we eliminated some sessions when family and patients werent fit, thus biasing the results toward positive.“ (Bericht, Anlage 1B).

Die E-mail von Prof. Birbaumer bezog sich genau auf die Situationen, die wir in dem Artikel auf S. 17, Zeilen 23-25, erwähnten: „Wenn eine Sitzung im Gang war, wurde sie für die Patienten F, G und W nie beendet. Bei Patient B wurde dreimal eine laufende Sitzung beendet, da Speichel abgesaugt werden musste und die Daten wurden in der Analyse nicht verwendet“.

Dort haben wir angegeben, dass wir Sitzungen ausschließen mussten, bei denen die Patienten wegen unmittelbarer Gesundheitsprobleme (z.B. drohendes Ersticken oder Absaugen von Speichel) am Experiment nicht teilnehmen konnten (auf der Basis der Information von Familienmitgliedern und den anwesenden Pflegekräften). Dies bezog sich natürlich nicht auf den Zustand eines Familienmitglieds.

In der Email v. 16.10.2017 schreibt Prof. Birbaumer weiter: „… 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, ...“ (Bericht, Anlage 1H, S. 1).

In der Anhörung am 25.02.2019 äußerte Prof. Birbaumer, dass ein solcher Ausschluss „manchmal“ vorgenommen worden sei. Befragt nach den Kriterien für einen solchen Ausschluss verwies er auf seine persönliche Beurteilungskompetenz.

c) In dem an die DFG gerichteten Schreiben der Betroffenen v. 17.05.2019, das Prof. Birbaumer dem Vorsitzenden der Kommission mit Email v. 18.05.2019 mitgeteilt hat, wird (unter 4. a.E.) schließlich folgendes dargelegt: „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 ...“ Damit wird eingeräumt, dass über die gerade angeführten Fälle des Gesundheitszustands der Patienten hinaus „sessions“ aufgrund der Anwendung eines (unklaren) Kriteriums „Zustand der Patienten erlaubt Differenzierung von Ja- und Nein-Antwort nicht“ ausgeschlossen wurden.

Wir können diese Beurteilung nicht nachvollziehen. Wir haben in dem Artikel auf S. 19, Zeilen 3-6 klar dargelegt, dass wir nur Sitzungen eingeschlossen haben, die die zufällige ja-nein-Differenzierung auf der Basis des NIRS-Signals überschritten („Wenn die Klassifikations-genauigkeiten mindestens drei aufeinander folgende Trainingssitzungen lang mit Fragen mit bekannten Antworten größer waren als die Schwelle, die wir für das Wahrscheinlichkeitsniveau bestimmt hatten, wurde ein neues Modell generiert, in dem wir die relative Veränderung in O2Hb über die drei Trainingssitzungen verwendeten, um online-Feedback zu geben“). Wir verweisen auch auf den Abschnitt BCI-Effectiveness auf S. 18, wo wir beschreiben, wie das Wahrscheinlichkeitsniveau definiert wurde.

Wir möchten noch einmal darauf hinweisen, dass es sich hier nicht um Laborexperimente handelt, sondern um Messungen im Heim der Patienten mit vielen Arten von Schwierigkeiten, die man dort erlebt, wenn Daten aufgenommen werden und die man berücksichtigen muss, da man ansonsten nicht-valide Daten sammeln würde. Da die Kommission und der Experte diese Datensammlung im Heim der Patienten nie sahen, glauben wir, dass es sehr schwierig für die Kommission war, korrekt zu evaluieren, wie die Daten gesammelt und analysiert wurden. Das gesamte Experiment war Teil eines Koselleck-Projekts des Beschuldigten Birbaumer, wo risikoreiche Experimente, die an die Grenzen des Erforschbaren gehen, explizit gefordert werden.

d) In dem genannten Schreiben wird weiter (unter 4.) eingeräumt, dass es bei der Datenerhebung technische Probleme gab, die häufig zum Ausschluss von „sessions“ führten („There were many instances when there was an error in the online data transfer ...“). Die Betroffenen legen dann dar: „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.“ Die von den Betroffenen zum Download bereitgestellte Datei „Readme_SessionDetails.docx“ enthält eine Liste der ausgeschlossenen „sessions“. Es waren bei Patient F 10, bei Patient G 7, bei Patient B 2 und bei Patient W 1 „session“, insgesamt 20. Dies steht in direktem Widerspruch zu der Aussage im Artikel, S. 9: „None of the sessions were eliminated in the analysis, ...“

Diese Sitzungen, die der DFG zur Überprüfung übergeben wurden, sind Sitzungen in denen Hardware- und Software-Interaktionsprobleme auftraten, die zu Daten führten, die man nicht analysieren konnte und somit die Sitzung wiederholt werden musste. Dies hatte mit Trigger-Problemen zu tun, die es unmöglich machten, die Daten zu analysieren, wie man in den Daten selbst auch sehen kann. Zu keinem Zeitpunkt wurden wegen der nicht korrekt übertragenen Trigger unerwünschte Daten ausgeschlossen. Die Triggerprobleme machten es de facto unmöglich, die Daten zu

analysieren. Wir haben zu keinem Zeitpunkt unerwünschte Daten ausgeschlossen, da der Fehler in dem gerätebasierten Datentransfer eine Analyse der Daten unmöglich machte. Dies kann man auch jederzeit anhand der zusätzlichen Daten, die wir der DFG übergeben haben, überprüfen, die auch der Kommission vorlagen. Da wir dies als einen Fehler im Gerät betrachtet haben, haben wir diese Sitzungen nicht berichtet, so wie sie normalerweise auch in anderen Experimenten nicht berichtet werden und auch keine Daten gesammelt wurden, die evaluiert werden konnten. Z.B. werden bei Experimenten im Magnetresonanztomographen Sitzungen mit Triggerfehlern einfach wiederholt, nachdem man den Fehler korrigiert hat. Genauso haben wir diese Sitzungen behandelt. Falls die Kommission an diesem Vorgehen Zweifel hat und wir fälschlicherweise diese Daten nicht hochgeladen haben, würden wir diese Sitzungen natürlich zu der hochgeladenen Datenbasis hinzufügen. Wir würden jedoch gerne noch einmal feststellen, dass PLoS Biology nur das Hochladen von Daten verlangt, die tatsächlich für die Ergebnisse relevant sind. Da diese Daten überhaupt nicht analysiert werden konnten und in den Ergebnissen nicht eingeschlossen wurden, gab es unserer Meinung nach keine Notwendigkeit, sie hochzuladen. Wir haben diese Frage auch in einem Brief an die DFG vom 17. Mai 2019 folgendermaßen beantwortet: „Alle Daten der Sitzungen mit Feedback und offenen Fragen wurden in die Datenanalyse und somit in die Publikation eingeschlossen. Wir schlossen Daten nur in der Modellbildungsphase (in den Trainingssitzungen) aus. Das Modell wurde mit den Trainingssitzungen gebildet, in denen die Differenzierung zwischen ja- und nein-Antworten 65% überstieg (siehe S. 19, Zeilen 3-7 des Artikels), wie wir es auch im anhängenden Excelfile für jeden Patienten beschrieben haben. Im Jahr 2014 wurde während des Experiments ein online-Datentransfer zwischen der Datenerhebung mit dem Gerät und dem BCI-Software Laptop durchgeführt. Es gab viele Fälle, in denen es zu einem Fehler im Online-Datentransfer zwischen den Laptops kam und dies führte zum Verlust von Datenpaketen und damit auch zum Verlust von den Markern der Trigger. Wir haben diese Dateien im angehängten Excelfile mit „Daten von dieser Sitzung wurden nicht analysiert, weil es ein online-Datentransferproblem gab“, markiert. Diese Daten wurden somit erhoben und gespeichert, aber wegen dieser Fehler nicht weiterverarbeitet. Dies bedeutet, dass wir alle Daten in den Sitzungen mit Feedback und offenen Fragen eingeschlossen haben und nur Daten in der Phase der Modellbildung, wo der Zustand des Patienten eine Differenzierung von Ja-Nein-Zuständen nicht erlaubte oder in Sitzungen, in denen der Trigger nicht funktionierte und deswegen keine Modellbildung erlaubte, ausgeschlossen haben. Dies ist auch in den Readme-Files dargestellt.“

Und

„In Tabelle 1 der Originalpublikation führten wir die Gesamtzahl der Trainings-, Feedback- und offene Fragen Sitzungen pro Patient auf. In der neuen Excel-Datei, die wir mit den Daten übermitteln, haben wir auch die Trainingssitzungen, die Datentransferprobleme hatten, wie wir sie oben beschrieben hatten, eingeschlossen (diese wurden nicht in die PLoS Biology Publikation eingeschlossen, da sie wegen der Fehler im online-Datentransfer nicht analysiert wurden, wie wir in Punkt 4 beschrieben haben und sie wurden deshalb nicht für PLoS Biology hochgeladen, jedoch für die DFG). Dasselbe bezieht sich auf die Abbildungen. Tabelle 1 gibt die komplette Anzahl der Sitzungen wieder, die analysiert wurden. Hier schlossen wir alle die Sitzungen ein, die durchgeführt wurden, d.h. auch diese, bei denen eine ja-nein-Antwort unter dem Zufallsniveau gegeben wurde, aber keine Sitzungen mit falschen Triggern. Die Sitzungen mit falschen Triggern wurden nur an die DFG übermittelt, da sie auf

Geräteproblemen basierten, jedoch wollten wir komplett in der Übergabe sogar der ausgeschlossenen Daten sein.“

e) Die Betroffenen haben also “sessions“ ausgeschlossen wegen
- Gesundheitszustand der Patienten (häufig)
- technischer Probleme (häufig)
- „Zustand des Patienten erlaubt Differenzierung von Ja- und Nein-Antwort nicht“.

Ein Ausschluss bestimmter „sessions“ und damit der darin ermittelten Daten ist nicht als solcher unzulässig. Jedoch müssen die Auswahlkriterien für einen Ausschluss vor Beginn der Datenerhebung definiert und dokumentiert werden. Außerdem muss gegenüber den Lesern des Artikels eine Offenlegung hinsichtlich der Anzahl der eliminierten „sessions“ sowie über die angewandten Kriterien und den Entscheidungsprozess bezüglich der Elimination erfolgen. Die ist hier nicht erfolgt. Weder im Artikel, S. 8 („Slow EEG Rythms‘ Relationship with fNIRS Classification Accuracy“) noch im „Response to: ‚Questioning ...‘“, S. 3-4 („Slowing of EEG and consciousness“) wird dargelegt, wie viele Sitzungen aufgrund welcher Kriterien ausgeschlossen wurden. Vielmehr wird im Artikel, S. 9 bewusst falsch behauptet, dass keine der „sessions“ von der Analyse ausgeschlossen worden sei. Insbesondere werden auch die nunmehr aufgeführten technischen Probleme nicht angesprochen. Auch aus den übergebenen NIRS-Daten lässt sich nicht erschließen, welche „sessions“ aufgrund welcher Kriterien ausgeschlossen wurden. Im Mai 2019 ex post erstellte Tabellen über den Ausschluss von Daten aufgrund bestimmter behaupteter Faktoren, die überdies weiter den Ausschluss aufgrund „state of the patient did not permit differentiation of yes, no states“ ohne klare weitere Angaben in den Raum stellen, ändern daran nichts.

Das beschriebene Vorgehen stellt also ein wissenschaftliches Fehlverhalten nach § 1 Abs. 2 Nr. 1. b) (Verfälschung von Daten durch Zurückweisen unerwünschter Ergebnisse ohne Offenlegung) dar.

Wir können diesem Argument nicht folgen. Wie wir in dem Artikel S. 17, Zeilen 23-25, festgestellt haben: „Wenn eine Sitzung im Gang war, wurde sie für die Patienten F, G und W nie beendet. Bei Patient B wurde dreimal eine laufende Sitzung beendet, da Speichel abgesaugt werden musste und die Daten wurden in der Analyse nicht verwendet. Der Gesundheitszustand der Patienten verlangte einen Ausschluss von Sitzungen manchmal in der Modellbildungsphase.

Die Unsicherheit über den Zustand des Patienten, wie man ihn aus dem Mangel an ja-nein-Differenzierung schließen konnte, führte dazu, dass einige Sitzungen ausgeschlossen wurden, wie wir auf S. 19, Zeilen 3-6 auch dargestellt haben: („Wenn die Klassifikations-Genauigkeiten mindestens drei aufeinander folgende Trainingssitzungen lang mit Fragen mit bekannten Antworten größer waren als die Schwelle, die wir für das Wahrscheinlichkeitsniveau bestimmt hatten, wurde ein neues Modell generiert, in dem wir die relative Veränderung in O2Hb über die drei Trainingssitzungen verwendeten, um online-Feedback zu geben“.)

Darüber hinaus kam es häufig auch zu technischen Problemen. Wir haben diese Sitzungen, in denen das Gerät nicht funktionierte und die wiederholt wurden, nicht beschrieben und in unserer Meinung ist deren Bericht auch nicht notwendig, weil diese Sitzungen nicht analysiert werden konnten und normalerweise auch nicht in der Literatur berichtet werden. Wir haben alle anderen Entscheidungen im Detail im Artikel berichtet, ebenso den Ausschluss von Daten auf der Basis von Nicht-

Differenzierung von ja-nein-Antworten, was sich auf die Trainingsphase bezieht, die als Basis für die Modellbildung diente. In anderen Worten, wir haben Sitzungen ausgeschlossen, um in der Lage zu sein, ein Modell für die ja-nein-Antworten in diesen sehr beeinträchtigen Patienten zu bilden, die kompromittierte Hirnaktivierungsmuster haben. Wir haben nie Sitzungen ausgeschlossen, wenn wir Feedback an die Patienten gegeben haben, also Sitzungen, die bei uns als Grundlage für die Bestimmung der Kommunikationsfähigkeit dienten.

Auf der Basis der sehr schwierigen Bedingungen im Heim der Patienten glauben wir, dass wir unser Bestes gegeben haben, um ein valides Modell und eine ausreichende Basis für einen Algorithmus zu finden, der bestimmt war, um später die Kommunikationsfähigkeit der Patienten zu erfassen. In den Feedbacksitzungen, wo wir die Kommunikationsfähigkeit beurteilten, haben wir niemals Sitzungen ausgeschlossen, sogar dann nicht, wenn wir glaubten, dass die Patienten vielleicht schlafen würden oder auf andere Weise nicht in der Lage waren, uns zu folgen, um die Daten nicht zu unseren Gunsten zu verfälschen.

Um es noch einmal zusammenzufassen: nach einer langen Trainingsphase, wo wir Sitzungen von der Analyse ausschlossen, um ein mathematisches Modell bilden zu können, das ja-nein-Antworten aus verschiedenen Hirnzuständen ablesen konnte, haben wir diese Sitzungen verwendet, die eine mehr als zufällige ja-nein-Differenzierung ergaben und haben aus diesen Sitzungen einen Classifier in einer bekannten und akzeptierten fünffachen Kreuzvalidierungsmethode gebildet. Wir haben dann diesen sogenannten Classifier verwendet, um die Kommunikationsfähigkeit der Patienten in einem neuen Satz von Sitzungen zu testen, den sogenannten Feedbacksitzungen, wo wir den Patienten Rückmeldung gaben und wo wir keine Sitzungen ausgeschlossen haben.

Daneben ist festzuhalten, dass Prof. Birbaumer sich bewusst war, dass sein Verhalten zu einer statistisch relevanten Verzerrung („bias“) in den Daten geführt hat, während im Artikel, S. 9 (vgl. o.) das Gegenteil behauptet wird. Außerdem wird im Artikel, S. 17 die Unterbrechung einer „session“ als „rare event“ qualifiziert, während Prof. Birbaumer mit Bezug auf die Jahre, in denen auch die hier in Frage stehenden „sessions“ stattfanden, beschreibt, er habe oft darauf gedrängt, eine „session“ auszuschließen („I pressed him often to eliminate a session“, vgl. o.).

Wir möchten nochmals darauf hinweisen, dass die Kommission missverstanden hat, was wir in den Trainingssitzungen für die Modellbildung gemacht haben und in den Feedback-Sitzungen, die zur Bestimmung der Kommunikationsfähigkeit verwendet wurden. Trainingssitzungen hatten den Zweck, das mathematische Modell zu bilden, während Feedback-Sitzungen die Basis für die Bestimmung der Kommunikationsfähigkeit waren. In der Trainingsphase verwendeten wir nur Sitzungen, in denen die Patienten Ja-Nein klar unterscheiden konnten und es ist sicherlich ein Bias in dem Sinne, dass wir versucht haben, die minimalen Kommunikationsfähigkeiten, die diese Patienten eventuell haben, zu maximieren, um den Algorithmus zu optimieren. Wenn wir im Zweifel waren, wurde eine Sitzung ausgeschlossen statt eingeschlossen, weil wir keine objektiven Mittel haben, um den Zustand des Patienten festzustellen. Dies war immer auf die Differenzierung des NIRS-Signals von Ja- und Nein-Antworten gerichtet. Dies unterscheidet sich von unterbrochenen Sitzungen in Bezug auf den Gesundheitszustand des

Patienten, die selten waren und die natürlich nicht in die Datenanalysephase eingingen.

So wie wir vorher festgestellt haben, haben wir in der Feedbackphase in der wir die Kommunikationsfähigkeit bestimmt haben und wo wir nie eine Sitzung ausgeschlossen haben, sogar wenn wir gedacht haben, dass der Patient nicht fit war, die Daten eher gegen uns gebiast. Dies haben wir auch auf S. 9, Zeilen 5-7 der Diskussion geschrieben, die sich auf die Feedbacksitzungen bezog: „In der Analyse haben wir keine der Sitzungen eliminiert und nur sehr wenige Sitzungen mussten wegen lebensrettender Maßnahmen unterbrochen werden: es gab somit keinen Bias zur Auswahl von „erfolgreichen Durchgängen, der die Ergebnisse verfälschen könnte“.

Schließlich lässt die folgende Aussage darauf schließen, dass Prof. Birbaumer sich selbst persönlich die Fähigkeit zuschrieb, vor Ort während der laufenden Sitzung mit den Patienten Ja- von Nein-Antworten zu unterscheiden: „ich bin zwar wie alle positiv gebiased, aber nicht so extrem, dass ich nicht ja von nein unterscheiden kann und dies in fast 100 Sitzungen mit mehreren Patienten!“ (Email von Prof. Birbaumer v. 16.11.2017, Bericht, Anlage 1L, S. 1).

Das stimmt in dem Sinn, dass ein erfahrener Forscher, der sich mit NIRS befasst, sehen kann, ob es eine Ja-Nein Differenzierung in den Daten auf der Basis des Hirnsignals gab. Wie wir in unserem Artikel geschrieben haben, war ein Ausschluss der Sitzungen jedoch immer auf einen dokumentierten Mangel von ja-nein-Differenzierung im Maschinenlern-Algorithmus, der verwendet wurde, basiert und nicht auf der Kommunikation zwischen den Experimentatoren während des Trainings. Die Kommission zitiert hier eine Aussage von Prof. Birbaumer außerhalb des Kontextes, in dem sie gemacht wurde. Prof. Birbaumer diskutierte in diesem Kontext die Rolle des physiologischen Signals im Vergleich zu dem maschinellen Lernalgorithmus und die theoretischen Grundlagen des Artikels, weil es eine Diskussion gibt, ob ein Maschinenlern-Algorithmus alleine wirklich die Nuancen des physiologischen Signals abbilden kann. Jedoch handelt es sich hier um ein allgemeines Problem und im PLoS Biology Paper wurde der Algorithmus des maschinellen Lernens so verwendet, wie er dort beschrieben wurde.

f) Ergänzender Hinweis

Die Email v. 16.10.2017 von Prof. Birbaumer (Bericht, Anlage 1G, S. 1) enthält folgende Aussage:

> „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 persormance [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.“

Auch hier wird zumindest dem ersten Anschein nach beschrieben, dass aufgrund persönlicher Entscheidung ohne weitere Kriterien Daten ausgeschlossen wurden („we eliminated ... trials“). Der beschriebene Vorgang steht jedoch in Zusammenhang mit Daten, die in die Publikation *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, eingegangen sind. Er war daher in die jetzige Untersuchung nicht einzubeziehen.

Dies wurde auch außerhalb des Kontexts zitiert und hat keinen Zusammenhang zum PLoS Artikel, da wir intensive Diskussionen über die wissenschaftlichen Hypothesen des Experiments hatten, d.h. wie die Form des physiologischen Signals mit der Klassifikation korreliert. Wie bereits oben betont, wurde, wenn es eine Diskrepanz gab, immer das Modell benutzt, nicht die visuelle Inspektion. Diese Diskussion fokussierte auf der Frage, ob physiologische Daten und die Klassifikation durch das Modell immer zusammenhängen sollten. Wie im Artikel dargestellt, war die visuelle Inspektion kein Kriterium für den Einschluss oder Ausschluss von Trainingssessions in das Modell. Wir finden es unverständlich, dass die Kommission Zitate aus wissenschaftlichen Diskussionen außerhalb des Kontextes verwendet, die mit dem Artikel nichts zu tun haben.

In Bezug auf diesen Artikel muss weiter darauf hingewiesen werden, dass die dort verwendeten Daten möglicherweise durch eine Fehlbedienung des benutzten NIRS-Geräts verfälscht wurden, so wie in der Email v. 20.10.2017 (Bericht, Anlage 1J, S. 1f.) dargestellt.

Prof. Birbaumer hat die Möglichkeit eines solchen Fehlers in einer Email v. 20.10.2017 bestätigt („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.“; Bericht, Anlage 1K, S. 1).

Die Kommission ignorierte unsere Antwort zu diesem Punkt an die DFG vom 22. April 2019, die wir auch der Kommission übergaben. Dort schrieben wir: „Wir fanden keinen Fehler in dem Artikel von Gallegos-Ayala von 2014 und glauben nicht, dass der Hinweisgeber irgendeinen Beleg für wissenschaftliches Fehlverhalten gegeben hat. Die Daten der Person, die vom Hinweisgeber erwähnt wurden und bei denen ein falscher Trigger existieren sollte, wurden nie in diesen Bericht eingeschlossen, da es sich nicht um einen Patienten, sondern um eine gesunde Pilotperson handelte. Der Hinweisgeber hatte keinen Zugang zu den Daten von Gallegos-Ayala, somit sind alle Vorwürfe, die er über diese Daten macht oder gemacht hat, nicht auf echte Daten basiert, sondern auf Annahmen und Vorwürfe.

Tatsächlich existiert die Art von Trigger, die er erwähnt, nicht in dem NIRS-Gerät, das wir in dem Artikel von 2014 verwendet haben (siehe die email von Gallegos-Ayala im Anhang Nr. 6 und dies ergibt sich auch aus der Gebrauchsanleitung des Geräts) und wir vermuten, dass diese Anschuldigungen nicht auf wissenschaftlichen Belegen basieren. Das NIRS-Equipment, das im PLoS Biology Artikel verwendet wurde, ist nicht dasselbe wie das, das in Gallegos-Ayala et al in 2014 verwendet wurde, sondern ein fortschrittlicheres Gerät, da alle Daten des PLoS Biology Artikels von 2014 an erhoben wurden. In dieser Art des Geräts gab es keine Trigger-Probleme, wie sie vom Hinweisgeber für die Daten im PLoS Biology Paper beschrieben wurden. Die Aussagen des Hinweisgebers bezogen sich entweder auf andere Tage und andere Experimente oder ein anderes NIRS-Gerät. Deswegen ist keine seiner Aussagen valide.”

2. Fehlende Offenlegung von Daten und Skripten

a) Der Artikel enthält Links zu verschiedenen Datensätzen. Darunter fehlt jedoch das Skript, mit welchem dem entsprechenden Auswertungsprogramm Vorgaben zur Auswertung der Daten gemacht werden.

Daneben enthält der Artikel auch keine Links zu denjenigen Daten, mit denen sich nachvollziehen ließe, dass das beschriebene Modell in Echtzeitsitzungen eine statistisch valide Zuordnung der jeweiligen Muster der Hirnaktivität eines Patienten zu einer (vom Patienten gedachten) „Ja"- oder „Nein"-Antwort erlaubt. Zwar sind Links vorhanden („S4 Table – S11 Table"), die sich jeweils auf Trainings- und Feedback-Sessions der Patienten F, B, G und W beziehen. Diese verweisen aber nur auf Tabellen (im Format MatLab und Excel), die nichts anderes enthalten als die in den Grafiken des Artikels dargestellten Daten, also das Endergebnis.

b) aa) Dadurch hält die Publikation nicht die Richtlinien der Zeitschrift, in der sie veröffentlicht wurde, ein. Diese sind in einer „Data Availability Policy", die für alle Zeitschriften von PLOS gilt, enthalten. (https://journals.plos.org/plosbiology/s/data-availability). Dort heißt es:

> „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) Außerdem verstößt die Veröffentlichung dadurch gegen Empfehlung 7 der Kommission „Selbstkontrolle in der Wissenschaft", die in der Denkschrift „Vorschläge zur Sicherung guter wissenschaftlicher Praxis" (1998, ergänzt 2013) enthalten sind. Dort (S. 21f.) heißt es:

> „Experimente und numerische Rechnungen können nur reproduziert werden, wenn alle wichtigen Schritte nachvollziehbar sind. Dafür müssen sie aufgezeichnet werden. Jede Veröffentlichung, die auf Experimenten oder numerischen Simulationen beruht, enthält obligatorisch einen Abschnitt ‚Materialien und Methoden', der diese Aufzeichnungen so zusammenfasst, dass die Arbeiten an anderem Ort nachvollzogen werden können."

c) Die Betroffenen haben in den Anhörungen ausgesagt, sie seien vom Herausgeber nicht aufgefordert worden, weitere Daten zur Verfügung zu stellen. Dr. Chaudhary hat versichert, er werde auf Anfrage die Rohdaten anderen Wissenschaftlern zur Verfügung stellen. Er hat der Kommission auf Aufforderung Daten im Umfang von ca. 5,2 GB zur Verfügung gestellt. Es handelt sich um die NIRS-Daten zu den vier Patienten F, G, B und W, die jeweils in einzelne Ordner „visit [n]" und mit einem Datum bezeichnete Unterordner (z.B. „2014-06-10") eingeordnet sind, sowie zur Auswertung benutzte Skripte (vgl. o. I. a. E.).

Auch aus diesen Daten lassen sich die im Artikel gemachten Angaben nicht nachvollziehen, da die Informationen fehlen, mit welchen Fragen jeweils gearbeitet wurde und was das BCI-System jeweils ermittelt hat. Außerdem sind keine der im Artikel genannten EEG-Daten enthalten.

Mit Email v. 18.05.2019 hat Prof. Birbaumer einen Download-Link zu Verfügung gestellt (vgl. o. I. a. E.) mit dem auf Daten zugegriffen werden konnte, aus denen sich für einen Teil der Experimente zumindest grundsätzlich die Möglichkeit ergibt, die NIRS-Daten mit Sounddateien und Textdateien in Beziehung zu setzen: In der Tabelle „SessionDetails“ verweist einerseits der „Folder Identifier“ auf die NIRS-Daten und ermöglicht andererseits die Angabe unter „Session“ den Bezug zu „Feedback results“ (für „feedback sessions“, bei Patient B z.B. 4) und „Open question results“ (für „open question sessions“, bei Patient B z.B. 2). Bei diesen „results“ handelt es sich um jeweils zwei Tabellen, von denen eine auf Sounddateien verweist (die Fragen). Die andere enthält bei den „feedback-sessions“ die vom BCI-System ermittelte Antwort, bei den „open question sessions“ die Einschätzung über die Richtigkeit der ermittelten Antwort (vgl. dazu aber u. 4.).

Es handelt sich dabei nach den Aussagen der Betroffenen um zum Zeitpunkt des Schreibens an die DFG neu hergestellte Übersichten („we have now also added“; „new excel sheet“, Schreiben v. 17.05.2014 unter 2. und 5., vgl. auch das Datum aller Dateien). Es muss also davon ausgegangen werden, dass sie zum Zeitpunkt der Übergabe der NIRS-Daten an die Kommission noch gar nicht existierten.

Unabhängig davon fehlen weiterhin die Informationen zu den „training sessions“. Damit ist der Nachvollzug und die Überprüfung der während der „training-sessions“ vollzogenen Modellentwicklug und -auswahl weiterhin nicht möglich. Der Datenbestand ist also weiterhin unvollständig. Schließlich fehlen auch die im Artikel dargestellten EEG-Daten.

Die Betroffenen haben vollständige Daten also weder mit der Publikation selbst zur Verfügung gestellt noch der Kommission Ende März 2019 und auch nicht der DFG und der Kommission Mitte Mai 2019.

Wir stimmen mit dieser Aussage nicht überein und würden gerne betonen, dass wir alle Daten zu PLoS Biology hochgeladen haben, die von den Gutachtern und den Editoren verlangt wurden und der PLoS Biology Datenrichtlinie entsprechen, die auf keinen Fall verlangt, dass alle Daten zur Verfügung gestellt werden, wie das Zitat der Datenrichtlinie, die wir unten einkopiert haben, zeigt.

Zusätzlich haben wir, um den Kommissionen mit ihrer Evaluation zu helfen, detaillierte Information über die Datenerhebung und –analyse in einer viel ausführlicheren Art dargelegt, als normalerweise bei Zeitschriften wie PLoS Biology verlangt wird. Dieser Versuch, den Kommissionen der Universität und der DFG in ihrer Beurteilung zu helfen, sollte jetzt nicht gegen uns ausgelegt werden. Auch der Umstand, dass wir geschrieben haben, dass wir diese Information jetzt hinzugefügt haben, sollte nicht gegen uns ausgelegt werden. Wir haben für die Gutachter der DFG und der Universitätskommission zusätzliche Informationen zusammengestellt, die natürlich 2017 schon verfügbar waren, aber von den PLoS Biology Gutachtern oder den Herausgebern, die auch unsere hochgeladenen Daten begutachtet haben, nie verlangt worden waren.

Diese Datenrichtlinie wurde uns von PLoS Biology geschickt:

“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.”

(Hervorhebung von Passagen von uns hinzugefügt)

Wir möchten hinzufügen, dass wir im Artikel klar definiert haben, welche Trainingssitzungen ausgeschlossen wurden und der Experte hätte leicht herausfinden können, welche Sitzungen das waren da es explizit im Artikel aufgeschrieben war (“ein neues Modell wurde nur gebildet, wenn 3 Sitzungen nacheinander den Zufallswert überstiegen“).

d) Die Verfahrensordnung der Universität Tübingen enthält keine Vorschrift, die der o.g. Empfehlung 7 der Kommission „Selbstkontrolle in der Wissenschaft“ entspricht. Allerdings könnte die Pflicht, Rohdaten sowie Skripte zur Verfügung zu stellen, um anderen Wissenschaftlern eine Überprüfung zu ermöglichen, eine unabdingbare ungeschriebene Regel im Sinn von § 1 Abs. 1 der Verfahrensordnung sein.

aa) Die Verfahrensordnung stellt dort auf die „geschriebenen oder ungeschriebenen Regeln ... in einem bestimmten wissenschaftlichen Fach oder einer wissenschaftlichen Fachrichtung“ ab, also auf die tatsächliche Fachkultur. Die Kommission kann nicht feststellen, dass es im betroffenen Wissenschaftsbereich als unabdingbar angesehen wird, bereits zusammen mit der Publikation alle Rohdaten und Skripte zugänglich zu machen. Insbesondere kann im Bereich medizinischer Forschung eine sofortige und vollständige Zurverfügungstellung in der Veröffentlichung u. U. sogar unzulässig sein.

bb) Davon zu trennen ist die Frage, ob die unvollständige Herausgabe an die Kommission selbst ein Fehlverhalten darstellt. Die Kommission ist sich bewusst, dass sie als zur Aufklärung des Vorwurfs eines wissenschaftlichen Fehlverhaltens eingerichtetes Gremium in einer besonderen Rolle steht: Als Untersuchungsgremium steht sie den von einem Fehlverhaltensvorwurf Betroffenen innerhalb eines Konfliktes in der Rolle eines Entscheidungsorgans gegenüber. Es ist also darauf zu achten, dass nicht im Rahmen einer Untersuchung durch die Kommission ein Verhalten der Betroffenen zum Fehlverhalten wird, das es außerhalb der Untersuchung nicht wäre.

Im vorliegenden Fall hat aber Dr. Chaudhary im Rahmen seiner Anhörung ausdrücklich darauf verwiesen, dass er und Prof. Birbaumer jedem Interessierten, ausgenommen allein der Hinweisgeber, alles offenlegen würden. Dr. Chaudhary betonte mehrmals, dass sie völlige Transparenz herstellen wollen. Dadurch hat die Kommission in diesem Fall nicht mehr verlangt, als die Betroffenen von sich aus angeboten haben.

Die unvollständige Herausgabe der relevanten und zum Nachvollzug der im Artikel beschriebenen Experimente notwendigen Daten an die Kommission stellt daher ein wissenschaftliches Fehlverhalten nach § 1 Abs. 2 Nr. 1 a) (Verfälschung von Daten durch Unterdrücken von relevanten Belegen) dar.

Wir stimmen nicht mit der Evaluation der Kommission überein. Wir erstellten die in unserer Meinung beste Beschreibung der Daten und Methoden, als wir sie zu PLoS Biology hochgeladen haben und folgten exakt der Datenrichtlinie von PLoS Biology. Wir haben die EEG-Daten nicht hochgeladen, da sie nicht der Hauptgegenstand des Artikels waren und nur als nicht ausreichend beschrieben wurden, die Kommunikation mit diesen Patienten zu ermöglichen. Die EEG-Daten wurden auch nie von den Gutachtern oder den Herausgebern von PLoS Biology verlangt. Sie waren auch nicht Teil der Ergebnisse, auf die sich die fNIRS-Kommunikationsfähigkeit bezieht. Wir teilten alle Daten, die für den Artikel relevant sind, mit der Kommission.

Jedoch haben wir bei den Nachfragen der Mitglieder des Ausschusses der DFG, die sehr detailliert und klar waren, festgestellt, dass mehr Information sinnvoll sein könnten, um jedes Detail dieser Publikation zu verstehen und haben deshalb eine sehr detaillierte Excel-Liste für diesen Zweck erstellt.

Insgesamt haben wir mehr als zwei Jahre für diesen sehr komplexen Analyseprozess verwendet und ein Jahr zur Erhebung der Daten und wir haben versucht, eine knappe und doch verständliche Beschreibung der Daten und des Analyseprozesses zu geben. Wir haben auch erklärt, dass wir sehr gerne alle Fragen über diesen Prozess jedem beantworten würden. Für uns als Wissenschaftler, die sehr tief in der Analyse von BCI-Daten und den entsprechenden Prozeduren stecken, ist es nicht leicht zu beurteilen, wieviel Dokumentation und Erklärung Personen außerhalb unseres Gebiets brauchen, um die Datenerhebung, -analyse und den Replikationsprozess zu verstehen. Wir haben von diesem Vorgang gelernt, dass eine noch detailliertere Dokumentation in zukünftigen Publikationen wünschenswert sein könnte und wir können dies natürlich auch für die Publikation, die derzeit untersucht wird, ergänzen. Wir möchten nochmals feststellen, dass wir nichts davon absichtlich getan haben oder um Replikation zu verhindern. Im Gegenteil, wir sind außerordentlich an einer Replikation interessiert und tatsächlich hat Dr. Sudhir Pathak unsere Daten repliziert und diese Ergebnisse auch mit uns publiziert.

Wie oben festgestellt, haben insgesamt 8 Gutachter unserer 2 Publikationen, Original und Kommentar, und die Herausgeber von PLoS Biology die Beschreibung der Datenanalyse und die hochgeladenen Daten als ausreichend detailliert befunden. Die Daten, die in PLoS Biology dargestellt wurden, wurden somit von den Gutachtern wie auch den Herausgebern des Journals beurteilt und sie waren glücklich mit den Ergebnissen und den Daten und erbaten nie, dass wir weitere Modifikationen durchführen sollten. Darüber hinaus sieht die Datenrichtlinie von PLoS Biology nicht vor, dass alle Daten publiziert werden sollen, sondern nur zusammengefasste Daten, die für die Abbildungen und Ergebnisse relevant sind, wie man in ihrer Datenrichtlinie nachlesen kann, die wir im vorhergehenden Abschnitt bereits hineinkopiert hatten. Wir möchten gerne noch einmal feststellen, dass PLoS Biology kein komplettes Hochladen von Rohdaten verlangt, sondern nur von diesen Daten, die die Schlussfolgerungen unterstützen.

Darüber hinaus glauben wir, dass die Kommission alle Daten hatte, die notwendig sind, um die Ergebnisse zu reproduzieren. Es ist möglich, dass der Sachverständige, der zugezogen wurde, nicht jeden Aspekt der Dokumentation und Analyse verstanden hat. Wir wären sehr glücklich gewesen, zusätzliche Informationen zur Verfügung zu stellen, hätten der Experte oder die Kommission uns danach gefragt.

3. Fehlende Daten

a) NIRS-Daten: Fehlende Tage mit Sitzungen (Patient B)

Im Artikel werden Angaben zu den Tagen gemacht, an denen mit den Patienten „training or feedback and open question sessions“ durchgeführt wurden (Artikel, S. 11, Table 2 mit Fn. 3). Daraus ergibt sich, dass bei Patient F an 14 Tagen, bei Patient G an 17 Tagen, bei Patient B an 12 Tagen und bei Patient W an 6 Tagen Sitzungen mit NIRS-Messungen durchgeführt wurden. Dies entspricht den Grafiken auf S. 7-10, in denen Ergebnisse aus 14, 17, 12, und 6 Tagen für die genannten Patienten dargestellt werden (vgl. auch Artikel, S. 20: „The number of days for each patient were: F, 14; G, 17; B, 12; and W, 6.“). Im Artikel, S. 18 heißt es: „The fNIR

S data was acquired online throughout all the sessions, namely trai- ning, online feedback, and open questions sessions.“

In den übermittelten Daten (vgl. o. 2. c)) befinden sich Ordner mit Ergebnissen aus folgenden Tagen (linke Spalte), hier gegenübergestellt mit den im Artikel genannten Tagen (rechte Spalte):

Patient F			In Artikel genannte Tage
	visit 1	24.-28.03.2014: 5 Tage	
	visit 2	15.-20.05.2014: 6 Tage	
	visit 3	04.-07.08.2014: 4 Tage	
	visit 4	04.-07.11.2014: 4 Tage	
		Summe: **19** Tage	*14*
Patient G			

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

In keinem Fall stimmt die Anzahl der Tage, zu denen Daten vorliegen, mit der Anzahl der Tage, für die im Artikel Auswertungen dargestellt werden, überein. Dadurch ist die Aussage im Artikel, S. 11, Table 2, Fn. 3 falsch. Es wurden bei den Patienten F, G und W mehr Sitzungen durchgeführt, aber nicht in die Analyse aufgenommen. Warum diese Sitzungen ausgeschlossen wurden, lässt sich nicht nachvollziehen.

Insbesondere werden aber zu Patient B Ergebnisse für *12* Tage angegeben, es liegen jedoch nur Daten für **8** Tage vor. Es werden also Ergebnisse für Tage dargestellt, für die keine Daten vorliegen.

Dies stellt ein wissenschaftliches Fehlverhalten nach § 1 Abs. 2 Nr. 1 a) (Erfinden von Daten) dar.

Wir haben die Daten, die in der linken Spalte dargeboten sind, nie gesehen. Als wir untersuchten, wie diese Daten zustande gekommen sein könnten, haben wir alle Stadien unserer Datendokumentation und des Datentransfers geprüft. Wir haben festgestellt, dass es ein Datentransferproblem in der Übergabe der Daten an die Universitätskommission am 2. April 2019 gegeben hat, was dazu geführt hat, dass einige zusätzliche Daten transferiert wurden und wenige andere Daten nicht übertragen wurden. Spezifisch wurde bei Patient F, Besuch 4, 04.11.2014-07.11.2014, eine spätere Sitzung, die nicht in der Publikation eingeschlossen wurde, fälschlicherweise der Kommission übergeben. Bei Patient B, wurde Besuch 3, 04.08.-07.08.2014 in die Publikation eingeschlossen, aber fälschlicherweise nicht der Kommission übergeben. Bei Patient W wurde Besuch 2, 15.12.-19.12.2014, eine spätere Sitzung, die in der Publikation nicht eingeschlossen wurde, fälschlicherweise der Kommission übergeben. Wir haben diesen Fehler damals nicht bemerkt. Wir haben jedoch die korrekte Anzahl und Sequenz der Sitzungen in dem Excelsheet im Brief an die DFG festgehalten, den wir am 18. Mai 2019 an die Kommission übergeben haben.

Im Brief der Kommission vom 22. Mai 2019 stellte Dr. Forster fest, dass die neuen Dateninformationen heruntergeladen worden wären und diskutiert worden seien. An diesem Punkt hätte die Kommission merken können, dass die Daten vom 02. April und vom 18. Mai nicht übereingestimmt haben. Sogar vorher hatte die Kommission

bereits festgestellt, dass die Daten, die wir am 02. April übertragen haben, nicht mit denen des Artikels übereingestimmt haben. Wir entschuldigen uns für dieses Versehen, aber wir haben erst jetzt festgestellt, dass die übertragenen Daten nicht den hochgeladenen Daten entsprechen.

Dr. Chaudhary glaubt, dass es einen Fehler in der Zusammenstellung der Daten für den Datentransfer gegeben hat, aber er kann die genaue Ursache für die teilweise falsche Übertragung nicht rekonstruieren. Nichtsdestotrotz hätten wir erwartet, dass die Kommission oder der Experte uns über diese Diskrepanz befragt, die man hätte erkennen können und die auch behoben hätte werden können. Die Kommission lud die relevanten Excel-Dateien am 20. Mai herunter und schrieb uns am 22. Mai, dass es keine weiteren Fragen gäbe. Wir vermuten deshalb, dass die Kommission die Diskrepanz in den Daten zu diesem Zeitpunkt übersah, jedoch war die Diskrepanz zu den Daten im publizierten Artikel bereits früher bekannt.

Im Übrigen ergibt sich hieraus auch, dass die Aussage über die zeitliche Durchführung der Experimente im Artikel, S. 17 („Each patient was visited 4 to 5 d in a month, except patient W.“) unzutreffend ist. Die Patienten F, G und B wurden zwischen 3 und 7 Tage pro Monat besucht.

Dies ist zutreffend, es hätte „im Durchschnitt“ heißen sollen. Wir entschuldigen uns für diesen Irrtum.

b) NIRS-Daten: Fehlende Unterordner (Patient F)

Die der Kommission übermittelten NIRS-Daten (vgl. o. 2. c)) zu den vier Patienten F, G, B und W sind jeweils in einzelne Ordner „visit [n]“ und mit einem Datum bezeichnete Unterordner (z.B. „2014-06-10“) eingeordnet. In diesen Unterordnern befinden sich wiederum solche mit dem Datum und einer Nummerierung, darin dann regelmäßig die Unterordner „Conditions“, „Detectors“ sowie „nirsSPM“ (mit Unterordner „nirs_data“). Für Patient F stellt sich die regelmäßige Verzeichnisstruktur ausschnittsweise wie folgt dar:

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

Dabei enthält der jeweilige Ordner (z.B. 2014-03-24_002) mehrere Dateien, darunter drei mit ca. 3 MB. Die Unterordner enthalten jeweils kleinere Dateien; der Unterordner „nirsSPM“ enthält nur einen Unterordner, dieser Unterordner „nirs_data“ enthält jeweils drei Dateien mit jeweils ca. 4 MB; so z.B.:

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

Diese Struktur wird allerdings nicht konsistent eingehalten. So fehlen mehrmals Ordner mit laufenden Nummern (z.B. Patient F, 2014-03-24: vorhanden sind 002, 003, 004, 005, 009; nicht vorhanden sind 001 sowie 006-008). Mehrmals fehlt der Ordner „nirsSPM\nirs_data“ (so bei Patient F, 2014-03-25_004, 2014-03-26_001 und 002 sowie 2014-03-27_001 und 002; bei Patient B, 2014-06-10_001 und 006).

Bei Patient F fehlt der Ordner „nirsSPM\nirs_data“ für die Tage vom 15.05.2014 bis 20.05.2014 (6 Tage) sowie für den 06.11.2014 (1 Tag) völlig. Im erstgenannten Zeitraum gibt es keinerlei Unterordner, für den zweitgenannten Tag nur den Unterordner „Detectors“.

Die fehlenden Sitzungen waren Zeitspannen, in denen das Gerät (NIRS-Gerät) getestet wurde. Auch Testsitzungen werden vom NIRS-Gerät kodiert, sogar wenn keine BCI-Sitzungen getätigt wurden. Zu den fehlenden SPM/NIRS-Daten: diese beziehen sich auf die vorverarbeitenden Daten, die nicht jedes Mal eingeschlossen wurden, aber dies sollte keine Konsequenzen haben, da es sich um eine spätere Verarbeitungsstufe handelt, die für die Dokumentation nicht erforderlich ist. Dasselbe gilt für den Ordner Detectors. Wir sind überrascht, dass die Kommission uns nie nach diesen Sitzungen gefragt hat, da wir diese natürlich sehr leicht zur Verfügung hätten stellen können.

Damit fehlen innerhalb der 19 Tage, an denen mit Patient F Sitzungen durchgeführt wurden, für 7 Tage NIRS-Daten. Es liegen also für **12** Tage NIRS-Daten vor. Im Artikel werden jedoch Ergebnisse für *14* Tage angegeben. Es werden also Ergebnisse für Tage dargestellt, für die keine Daten vorliegen.

Dieser Unterschied bezieht sich auf die fälschliche Datenübertragung vom 02. April 2019, die richtigen Daten und die Daten im Artikel stimmen 100% überein und wir hätten diese Diskrepanz aufklären können, wenn die Kommission uns darüber befragt hätte.

Dies stellt ein wissenschaftliches Fehlverhalten nach § 1 Abs. 2 Nr. 1 a) (Erfinden von Daten) dar.

Wir weisen diese Bewertung zurück. Der Unterschied bezieht sich auf die fälschliche Datenübertragung vom 02. April. Die richtigen Daten und die Daten im Artikel stimmen 100% überein und wir hätten diese Diskrepanz aufklären können, wenn die Kommission uns darüber befragt hätte. Darüber hinaus enthielt die Datei, die die Kommission am 20. Mai herunterlud, die richtigen Sitzungen.

Besondere Relevanz erhält dieser Befund dadurch, dass im Zeitraum vom 15.05.2014 bis 20.05.2014 nach der im Mai 2019 zur Verfügung gestellten Übersicht („PatientFSession-Details“) 4 (von 6) „feedback sessions“ und 1 (von 3 nach Artikel, S. 7, Fig. 2 bzw. 2 nach S. 6, Table 1) „open question sessions“ stattgefunden haben sollen. Selbst wenn man die ex post erstellte Zuordnung der verschiedenen Daten und Informationen akzeptieren wollte, werden also im Artikel Ergebnisse dargestellt, für die keine NIRS-Daten vorliegen.

Im Übrigen zeigt der vorhandene Datenbestand, dass die Aussage im Artikel, S. 17 („Three to four sessions were performed each day, ...“) unzutreffend ist. So wurden am 24.03.2014 bei Patient F 5 Sitzungen, am 15.08.2014 bei Patient B 5 Sitzungen, am

21.06.2014 bei Patient G 8 Sitzungen sowie am 16.12.2014 bei Patient W 5 Sitzungen durchgeführt.

Einige dieser Daten waren im Artikel nicht inkludiert, jedoch fälschlicherweise im April übertragen worden, wie wir oben bereits dargestellt haben.

c) NIRS-Daten: Fehlende Zuordnung

aa) Die übergebenen NIRS-Daten enthalten insgesamt keine Informationen darüber, welche NIRS-Daten sich auf welche Schritte (training, feedback, open questions) in den im Artikel beschriebenen Experimenten beziehen. Die übergebenen Daten enthalten auch keine Informationen darüber, welche Fragen jeweils in einer Sitzung gestellt wurden und welche Antworten identifiziert wurden. Schon daher lässt sich nicht nachvollziehen, welche Daten überhaupt auf welche der im Artikel dargestellten Schritte in den Experimenten zu beziehen sind. Entsprechend lässt sich aus den übergebenen Daten nicht nachvollziehen, wie die im Artikel beschriebenen Ergebnisse erzielt wurden. Das bedeutet, dass die im Artikel beschriebenen Ergebnisse nicht in nachvollziehbarer Weise mit Daten belegt sind.

Dies gilt insbesondere für die Daten zu den sog. „open question-sessions“ (zu deren Anzahl s. Artikel, S. 6, Table 1 bzw. S. 7 Fig. 2). Bei diesen Sitzungen wurden nach der Darstellung im Artikel sog. offene Fragen gestellt und eine Ja- oder Nein-Antwort identifiziert. Es müssten also Angaben (in Form eines Textes) vorliegen, aus denen sich ergibt, welche Fragen gestellt wurden, welche Antwort jeweils identifiziert wurde und welche Daten dem jeweiligen Identifikationsvorgang hinsichtlich einer bestimmten Frage zuzuordnen sind. Erst daraus ergäbe sich die Eigenschaft eines oder mehrerer Datensätze, Daten einer „open question-session“ zu enthalten.

Die übergebenen Daten enthalten aber keinerlei entsprechende Informationen. Selbst wenn man (fälschlicherweise) zugestehen würde, dass die ersten beiden Schritte in den Experimenten der Entwicklung und dem Test eines Modells dienten und daher hierzu nicht die entsprechenden Fragen und identifizierten Antworten vorliegen müssten, müssten dennoch die Daten aus den sog. „open question-sessions“ eindeutig mit entsprechenden Fragen und identifizierten Antworten verbunden werden können. Dies ist nicht der Fall. Es liegen in den übergebenen Daten keine als solche identifizierbare Daten zu „open question-sessions“ vor. Es werden also Ergebnisse dargestellt, zu denen keine Daten vorliegen.

Dies stellt ein wissenschaftliches Fehlverhalten nach § 1 Abs. 2 Nr. 1 a) (Erfinden von Daten) dar.

Wir stimmen der Einschätzung, dass hier wissenschaftliches Fehlverhalten stattgefunden hätte, nicht zu. Wir haben keine Daten erfunden. Wir stimmen zu, dass wir noch extensivere Dokumentationen der Daten einschließlich der Textdateien hätten übermitteln können. Jedoch wurde dies nie von den Gutachtern der Zeitschrift verlangt. Wir haben diese Daten der Kommission bereits übergeben und sehen kein Problem darin, diese Daten als zusätzliche Information zum Artikel ebenfalls hochzuladen. Wie vorher festgestellt, sind in der BCI-Forschung solch detaillierte Datendokumentationen bislang nicht die Regel gewesen und wir haben nicht absichtlich diese sehr detaillierte Dokumentation ausgelassen.

Wir hätten gerne mehr Daten zu den offenen Fragen zur Verfügung gestellt, wenn die

Kommission uns danach gefragt hätte, jedoch erhielten wir nie zusätzliche Anfragen.

bb) Die im Mai 2019 erstellten und übermittelten Verknüpfungen der NIRS-Daten mit den weiteren Informationen über Tabellen (vgl. o. 2. c) ändern hieran nichts.

Zum einen räumen die Betroffenen in ihrem Schreiben an die DFG v. 17.05.2019 (dort unter 2.), das dem Vorsitzenden der Kommission durch Prof. Birbaumer mit Email v. 18.05.2019 mitgeteilt wurde, ausdrücklich ein, dass

- für alle „feedback sessions“ mit Patient G sowie
- für die „feedback sessions“ mit Patient F innerhalb von „visit 3“ (04.-07.8.2014) keine Aufzeichnungen über die jeweiligen Ja- oder Nein-Antworten vorliegen.

Dies ist nicht zutreffend; wir haben festgestellt, dass für diese Sitzung computergenerierte Übersichtsdateien gebildet wurden, die von uns nicht manipuliert werden konnten. Tatsächlich war es so, dass die Prozentzahl der Antworten, die die Schwelle pro Patient überstieg, in diesen Übersichtsdateien niedriger war als in den individuellen Dateien, d.h. dass hier die Daten sogar gegen unsere Hypothese der Kommunikationsfähigkeit gebiast waren.

In der entsprechenden ReadMe-Datei in den neuen Daten heißt es für 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.“ Die in den genannten Schreiben gemachten Ausführungen über die Ermittlung der Prozentzahl richtiger

Antworten in diesen „sessions“ sind unklar. Jedenfalls kann die Ermittlung nicht mehr nachvollzogen werden. Im Artikel werden also auch nach den eigenen Ausführungen der Betroffenen Ergebnisse dargestellt, zu denen keine Daten vorliegen.

Dies ist nicht zutreffend. Im Brief an die DFG vom 17. Mai 2019 haben wir ganz klar gesagt: „Die entwickelte BCI-Software hatte auch eine Vorkehrung, in der der Prozentsatz der korrekten Antworten berechnet wurde. Deshalb haben wir für Sitzungen, in denen das System die individuellen Antworten nicht gespeichert hat, den Gesamtprozentsatz der korrekten Antworten gespeichert und später diesen Prozentsatz als das Ergebnis der jeweiligen Sitzung verwendet.“ Somit hatte die Kommission die Information, wie diese Durchgänge berechnet wurden.

Für Patient F ergibt sich damit i.ü., dass zu den drei „open question sessions“

- für eine keine NIRS-Daten (vgl. o. 3. b) a. E.) und
- für die beiden anderen keine Aufzeichnungen vorhanden sind.

Es fehlen weiterhin vollständig die Informationen zu den „training sessions“.

Das ist alles nicht zutreffend. Die Kommission hätte feststellen können, dass es eine Diskrepanz zwischen der Datenübertragung vom 02. April 2019 und den heruntergeladenen Daten vom Mai 2019 gab und hätte uns darüber am 22. und am 29. Mai 2019 befragen können. Darüber hinaus hätte der Experte, der diese Diskrepanz bemerkt hat, uns auch schon viel früher um diese Information bitten können. Darüber hinaus war die Information zu den Trainingssitzungen im Artikel verfügbar.

Schließlich handelt es sich bei den Angaben vom Mai 2019 um eine ex post durchgeführte Rekonstruktion. Zum Zeitpunkt der Veröffentlichung des Artikels (Januar 2017) existierte diese Zuordnung als solche nicht und hätte nicht zur Verfügung gestellt werden können. Entsprechend hat auch Dr. Chaudhary Anfang April 2019 nur die NIRS-Daten und damit nur Datenmaterial ohne nachvollziehbaren Bezug zu den im Artikel dargestellten Experimenten und Ergebnissen übergeben.

Wir können diesen Aussagen nicht folgen. Wir haben noch die Zeit-markierten Dateien von 2017, die all diese Daten enthalten. Wir haben ein Excelfile für die DFG und die Kommission neu zusammengestellt, um es den Kommissionen leichter zu machen, die Daten zu verstehen.

cc) Selbst wenn man die im Mai 2019 neu erstellten Zuordnungen akzeptieren wollte, ergibt sich kein anderes Ergebnis.

Bei Patient B wurden zwei „open question sessions“ durchgeführt. Bzgl. der **ersten** „open question session“ sind vorhanden „V2D4b5_QuestionList.txt“ und „V2D4b5_result.txt“. „V2D4b5_QuestionList.txt“ wird durch folgenden Text eingeleitet:

> „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.“

Wir entschuldigen uns, wenn dies schwer nachzuvollziehen war, jedoch hätte man uns sehr leicht befragen und so das Problem klären können. An diesem Tag war das Programm für die offenen Fragen, das einen anderen Code hat als das Programm für die Fragen mit bekannten Antworten, nicht funktional. Man muss hier berücksichtigen, dass die Forschungsteams mehrere hundert km reisen, um mit diesen Patienten für einige Tage zu arbeiten. Das Problem konnte vor Ort nicht behoben werden. Das Team entschied deshalb, das sehr teure Experiment nicht zu unterbrechen und nach Hause zu fahren, sondern das Programm zu verwenden, das normalerweise für die Fragen mit bekannten Antworten verwendet wird und die Fragedateien für diesen Zweck umzubenennen, wie es oben beschrieben wurde. Unserer Meinung nach ist es ein absolut akzeptabler Vorgang, da es sich nur um eine formale Operation handelte und die Daten, die gesammelt wurden, in keinster Weise betroffen hat.

Um dies zu dokumentieren, haben wir auch eine Datei mit einer Zeitmarkierung zur Verfügung gestellt, in der sowohl die umbenannte wie auch die originale Datei vorhanden sind. Wir würden hier gerne noch einmal erklären, wie die Korrektheit der Fragen einer Feedbacksitzung und einer Sitzung mit offenen Fragen berechnet wird. Wir haben dies in den nachfolgenden Abschnitten zusammengefasst:

In jeder Sitzung (Training und Feedback) werden 20 bekannte Fragen dargeboten. Für diese wird eine Fragenliste gebildet, diese Frageliste hat 10 wahre und 10 falsche Sätze in randomisierter Reihenfolge. Die Antwort zu den falschen Sätzen ist 0 – nein und der Name des Satzes beginnt mit 001 (Nummer), während die Antwort für einen wahren Satz 1 ist, ja, und der Name des Satzes beginnt mit 002 (Nummer). Für Feedback-Sitzungen besteht die Ergebnis-Datei, die für jede Sitzung gebildet wird, aus der Antwort zu jeder Frage, die vom Classifier vorhergesagt wird als 0 (d.h. der Classifier hat ein nein vorhergesagt und das BCI-System sagt, ihre Antwort war nein) oder 1 (d.h. der Classifier hat ja vorhergesagt und das BCI-System hat geantwortet, ihre Antwort war ja). Um die % korrekten Antworten in einer Sitzung zu berechnen, wird die dargebotene Antwort, d.h. die Antwort der Frage in der Fragenlistedatei mit dem Label der vorhergesagten Antwort im Ergebnisfile abgeglichen, d.h. wenn das Label der dargebotenen Frage 1 und das vorhergesagte Label auch 1 ist, dann wird von einer korrekten Vorhersage ausgegangen, während, wenn das Label der Frage 0 ist und das vorhergesagte Label 1 ist, die Antwort falsch vorhergesagt wurde usw. Für die Sitzungen mit offenen Fragen wird, da wir die Antwort auf die Frage nicht wissen, die Frageliste nur mit einer einzigen Antwort (wir hatten hier die Nummer 2 als Antwort in unserem Programm verwendet) für alle Fragen verwendet und der Name des Satzes beginnt mit 003 (Nummer). Wie immer kreiert das BCI auch hier ein Ergebnisfile mit der vorhergesagten Antwort für jede Frage als 0, d.h. der Classifier hat nein vorhergesagt und das BCI hat gesagt, ihre Antwort war nein oder 1, d.h. der Classifier hat ja vorhergesagt und das BCI sagte, ihre Antwort war ja. Diese Antwort ist dann mit der Antwort abgeglichen worden, die die Familienangehörigen für die richtige halten, um die Korrektheit der Beantwortung der Fragen der Sitzung zu bestimmen.

Als wir ein Problem hatten, das Modul der offenen Fragen beim Patienten am Bett durchzuführen, haben wir die offenen Fragen als bekannte Fragen umbenannt, d.h. 003 (Nummer) als 002 Nummer und 001 (Nummer) in einer randomisierten

Reihenfolge, wo 10 von den 20 offenen Fragen mit 002 und die anderen 10 als 001 benannt wurden, um die offenen Fragen mit dem Code der bekannten Fragen zu präsentieren und den dysfunktionalen Code zu umgehen. Für die Analyse wurde die Antwort zu jeder offenen Frage mit der vermuteten Antwort des Familienmitglieds in Zusammenhang gebracht, um die gesamte Genauigkeit dieser Sitzung zu bestimmen.

Es ist zunächst festzuhalten, dass trotz eines Problems bei der Datenerhebung die Daten benutzt und im Nachhinein Dateien umbenannt wurden. Der zitierte Text soll wohl besagen, dass in dieser Datei (im Gegensatz zu der entsprechenden Übersicht bei „feedback sessions“) der Beginn des Dateinamens mit 001_ bzw. 002_ nicht „true sentence“ bzw. „false sentence“ bedeuten und der in der jeweils folgenden Zeile enthaltene Wert nicht jeweils 1 für true und 2 für false ist. (Vgl. bei Patient F, Datei 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.)

Wie oben dargestellt, kam es zu keiner Datenmanipulation.

Weiter enthält „V2D4b5_QuestionList.txt“ die Dateinamen von 20 Sounddateien. „V2D4b5_result.txt“ enthält jedoch *21* 0- bzw. 1-Einträge. Zu keiner der Fragen ist also eine eindeutige Zuordnung möglich, da der überzählige Wert an jeder Stelle der Aufzählung stehen kann.

Die 0 und 1 in diesem Ergebnisfile für die offenen Fragen beziehen sich auf die Antworten, die vom Classifier auf der Basis der Software für die bekannten Antworten bestimmt wurden und eine zusätzliche Antwort wurde wegen eines Fehlers von der Software addiert. Da es sich hier Jedoch um eine Sitzung mit offenen Fragen handelte, waren die 0 und 1 Antworten für die Feedbacksitzung ohne Bedeutung und wurden nicht analysiert, da die Sitzungen mit offenen Fragen keine bekannte ja/nein-Antwort haben. Wir haben dies am Beginn des Ergebnistextfiles, das wir an die DFG am 17. Mai geschickt haben und das von der Kommission am 20. Mai heruntergeladen wurde, beschrieben. Wie wir dort geschrieben haben, hat der Inhalt dieser Datei keine wirkliche Bedeutung, da die Antwort zu den offenen Fragen mit der geschätzten Antwort des Familienmitglieds in Bezug gesetzt wurde und zwar in Echtzeit und diese Daten wurden dann eingetragen und benutzt.

„V2D4b5_result.txt“ wird mit folgendem Text eingeleitet: „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.“ Es folgen 14 1- und 7 0-Einträge. Daraus ergäbe sich eine „classification accuracy“ von 14/21, entspricht 66,67 % (falls man mit 20 Gesamtfällen rechnete, ergäben sich 70 %). Im Artikel, S. 9, fig. 4, werden deutlich höhere Werte dargestellt.

Die Klassifikationsgenauigkeit einer Sitzung mit offenen Fragen wird so berechnet, dass man das vorhergesagte Label jeder Antwort, das vom Classifier vorhergesagt wurde, entweder als 0, d.h. der Classifier hat ein nein vorhergesagt und das BCI sagt, ihre Antwort war nein oder 1, d.h. der Classifier hat ja vorhergesagt und das BCI sagt, ihre Antwort war nein mit der Antwort, die vom Familienmitglied für richtig erachtet wird, in Zusammenhang bringt. Wenn der Classifier 0 vorhersagt, was nein

bedeutet und das Familienmitglied auch glaubt, dass die Antwort nein ist, dann wurde dies als korrekt klassifizierte Antwort eingetragen usw. Das Matching von den Labels, über die die Mitglieder der Kommission sprechen, ist valide für eine Sitzung mit bekannten Antworten, aber das Mitglied der Kommission hat hier übersehen, dass es sich um eine Sitzung mit offenen Fragen handelte, die wie eine Feedbacksitzung behandelt wurde, was heißt, das das Label 0 und 1, das in Wirklichkeit 2 ist, keinerlei Bedeutung hat.

Wie in unserer Antwort oben bereits dargestellt, möchten wir hier nochmals festhalten, dass für die Sitzungen mit offenen Fragen, da wir die Antwort nicht wissen, nur ein einziges Label, nämlich 2 für alle Fragen verwendet wurde und dass der Name des Satzes mit 003 beginnt (Nummer). Das BCI kreiert hier ein Ergebnisfile mit der vorhergesagten Antwort für jede Frage als 0, z.B. wenn der Classifier nein vorhergesagt hat und das BCI sagt, ihre Antwort war nein oder 1, wenn der Classifier ja gesagt hat und das BCI, ihre Antwort war ja vorhergesagt hat. Diese Antwort wird dann mit der Antwort, die das Familienmitglied als korrekt annimmt, abgeglichen und aus diesen wird der Prozentsatz der richtigen Antworten für die Sitzung bestimmt.

Bzgl. Der **zweiten** „open question session“ bei Patient B wird in der Übersichtstabelle die „session“ V2D5b*5* angegeben. Vorhanden sind aber die Tabellen „V2D5b*6*_Question-List.txt“ und „V2D5b*6*_result.txt“. Es liegen also nicht die Informationen vor, um die entsprechenden Daten zuordnen zu können.

Als wir die Datei umbenannt haben, um sicherzugehen, dass die Gutachter bei der DFG in der Lage sein würden, die Dateien mit dem jeweiligen Fall exakt zusammenzubringen, haben wir unabsichtlich die Zahl vertauscht, wir entschuldigen uns für diesen Fehler. Der Dateiname hätte ...05 sein sollen.

„V2D5b6_result.txt“ enthält (nach dem gleichen Einleitungssatz) überdies 16 0- und 4 1-Einträge. Daraus ergäbe sich eine „classification accuracy“ von 4/20, entspricht 20 %. Im Artikel, S. 9, fig. 4, werden deutlich höhere Werte dargestellt.

Wie oben bereits dargestellt, handelt es sich hier um eine Sitzung mit offenen Fragen, die als eine Sitzung mit bekannten Fragen, also wie eine Feedbacksitzung verwendet wurde, nachdem wir die offenen Fragendateien umbenannt hatten. Dies ist der Dateiname, der für bekannte Fragen verwendet wurde und wir haben diese in einer randomisierten Reihenfolge umbenannt, um das dysfunktionale Modul für offene Fragen zu umgehen, das in der BCI-Software an diesem Tag vorhanden war. Die Genauigkeit der beantworteten Fragen wurde hier mit der Einschätzung der Familienmitglieder in Zusammenhang gebracht. Die im Artikel dargestellte Zahl ist deshalb korrekt.

Für die jeweiligen Tabellen zu den beiden „open question sessions“ bei Patient B fehlen schließlich die Informationen zu den vom BCI-System ermittelten Antworten, auf die sich die Richtigkeitseinschätzung auf Basis des Feedback der Familienangehörigen beziehen soll.

Wie oben dargestellt, wurde "wurde die Klassifikationsgenauigkeit bei einer Sitzung mit offenen Fragen so bestimmt, dass das durch den Classifier vorhergesagte Label jeder Antwort (entweder 0, d.h. der Classifier hat die Antwort als nein vorhegesagt und das BCI sagte „Deine Antwort war nein" oder 1, d.h. der Classifier sagte ja voraus und das BCI sagte, "Deine Antwort war ja" mit der Antwort des Familienmitgliedes abgeglichen wurde. Wenn der Classifier z.B. 0 für nein vorhersagte und das Familienmitglied auch "nein" als richtige Antwort einschätzte, dann ging man von einer korrekten Klassifikation aus usf".

Wie wir auf S. 13 unseres Manuskripts Chaudhary et al 2017 festgestellt haben, „wir müssen bei unserer Beurteilung der Antworten zu den offenen Fragen vorsichtig sein" und nochmals auf S. 18 unseres Manuskripts Chaudhary et al 2017, „Die Validität der Antworten auf die offenen Fragen kann nur über a) Augenscheinvalidität (z.B. Fragen zum Schmerz, wenn eine offene Wunde da ist); (b) Stabilität über die Zeit; (c) externe Validität, eingeschätzt durch Familienmitglieder und Pflegende; und (d) interne Validität zwischen Fragen (z.B. Konkordanz zwischen der Antwort zu „ich lebe gerne" mit der Antwort „ich fühle mich selten traurig" [die alle Patienten —außer W—regelmäßig erhielten] ermittelt werden.

Wir können somit zeigen, dass die Berechnung und die Bedeutung der Klassifikation der offenen Fragen in unserer Originalpublikation bereits erklärt wurde, die die Universitätskommission jedoch nicht beachtete und verwendete.

Insgesamt würden im dargestellten Fall selbst dann, wenn man die von den Betroffenen ex post erstellte Zuordnung der Daten und Informationen akzeptierte, im Artikel Ergebnisse dargestellt, zu denen keine Daten vorliegen.

Dies ist nicht zutreffend. Wir haben keine Dateien selbst erfunden, wir haben lediglich keine so detaillierte Information, wie von den Kommissionsmitgliedern im Nachhinein verlangt, zur Verfügung gestellt, aber wir können diese Information jederzeit beibringen, wenn erforderlich und hätten dies auf Nachfrage auch getan. Die Gutachter von PLoS Biology hielten dies nicht für notwendig.

Weitere Inkonsistenzen liegen darin, dass in den beiden QuestionList-Dateien auf Sounddateien vom gleichen Tag verwiesen wird, obwohl die betroffenen „sessions" an zwei verschiedenen Tagen stattfanden. Außerdem steht die Uhrzeit der darin aufgelisteten Sounddateien in Gegensatz zu deren Reihenfolge.

Alle Audiodateien wurden zu unterschiedlichen Zeitpunkten vor den Sitzungen erstellt. Dies hätten wir auch in einem Gespräch mit der Kommission erläutern können.

4. Mögliche Datenverfälschung durch fehlerhafte Analyse

a) In den übergebenen Daten finden sich zwei Skripte, „NIRs_trainmodel1.m" und

„trainSVMlinearclassifier.m“ (zu letzterem vgl. Bericht, Anlage 1A, S 10, Text v. 9.10.2017, Appendix A.2). „NIRs_trainmodel1.m“ lädt die NIRS-Daten und führt eine Vorverarbeitung der Rohdaten durch. Zumindest wenn die Variable „feature“ die Werte 2 oder 3 hat, wird in Zeilen 204 und 221 die Funktion „fea_csptrain“ aufgerufen, die, wahrscheinlich, die relevanten Features aus den Rohdaten extrahiert. Diese für die Ergebnisse (potentiell) extrem wichtige Funktion wurde der Kommission jedoch nicht zur Verfügung gestellt. In Zeile 227 ruft „NIRs_trainmodel1.m“ dann „trainSVMlinearclassifier.m” auf. Die durch die Anwendung dieses Skripts erhaltenen Ergebnisse werden durch „NIRs_trainmodel1.m“ nur noch abgespeichert (Zeilen 229-245).

Das Skript „trainSVMlinearclassifier.m“ führt eine Parameteroptimierung mit Kreuzvalidierung („cross-validation“) durch (Zeilen 12 und 15, flag „-v fold“). Dabei wird hinsichtlich der „classification accuracy“ ein kreuzvalidierter Vergleich zahlreicher Parameterkombinationen durchgeführt und die Kombination mit der höchsten „classification accuracy“ festgehalten. Für die beste Kombination von Hyperparametern „i“ und „k“ wird dann aber nochmals eine lineare „Support Vector Machine“ trainiert, jedoch ohne Kreuzvalidierung (Zeile 23). Ohne Anwendung der statistischen Methode der Kreuzvalidierung führt ein durch reine Optimierung erhaltenes Modell bei der Anwendung auf den gleichen Datenbestand, anhand dessen es optimiert worden ist, zu einer Verfälschung der Ergebnisse in Richtung auf signifikante Ergebnisse („bias“). Um nicht zu einer solchen Verfälschung zu führen, müsste es auf einen unabhängigen Datenbestand angewandt werden (das ist die Essenz der Kreuzvalidierung).

Dies lässt sich mit Hilfe der vorliegenden Skripte nicht abschließend prüfen: Würde das in Zeile 23 erhaltene Modell nur auf neue Online-Daten angewendet, wäre die Vorgehensweise korrekt. Sind die im Artikel berichteten Werte jedoch die, die durch die Optimierung in Zeile 23 erhalten wurden, dann sind die Werte mit hoher Wahrscheinlichkeit zu hoch („bias“). Die zur Beurteilung dieser Frage notwendigen Informationen und MATLAB-Skripte wurden der Kommission nicht zur Verfügung gestellt.

Die Kommission hat korrekt verstanden, dass das Modell auf einer Kreuzvalidierung basierte und dass die danach generierte Optimierung mit dem Support Vector Machine tatsächlich auf neue Daten angewendet wurde (Feedbacksitzungen). Wir würden gerne darauf hinweisen, dass das Hochladen dieser Arten von Skript bislang nicht der Normalfall in BCI-Artikeln war, (z.B. „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." Nature 485.7398 (2012): 372., und andere). Dies basiert auch auf Problemen mit dem Schutz geistigen Eigentums und es war überdies von PLoS Biology nicht vorgesehen. Wir hätten das Skript natürlich der Kommission zur Verfügung gestellt, wenn sie uns danach gefragt hätte.

b) Unabhängig davon ergibt sich jedoch aus dem Bericht der Vertrauenspersonen das Folgende: Die Betroffenen hatten spätestens mit der Email v. 09.10.2017 mit einem entsprechenden, detaillierten „Report“ (Bericht, Anlage 1A) Kenntnis von dieser Problematik („bias“). Die zeitlich darauf folgende Email-Kommunikation zeigt, dass die Betroffenen sie selbst zugestehen, Email v. 16.10.2017 (16:18 Uhr) von Prof. Birbaumer (Bericht, Anlage 1G, S. 1): „…: we used a hypothesis driven feature and model building, which is statistically not correct but physiologigcally [sic] plausible.“ Entsprechend heißt es in der Email von Prof. Birbaumer v. 16.10.2017, (16:54 Uhr) (Bericht, Anlage 1H, S. 1): „the model we built was based on our a priori hypothesis how it should look like ...“ In der Email von Prof. Birbaumer v. 16.11.2017, (Bericht, Anlage 1L, S. 1) heißtes:

> „Ja, wir arbeiten an der exakten Korrektur von Fehlern, ich möchte noch die unabhängigen Videoauswertungen abwarten, … Du hast völlig recht, Fehler muss man korrigieren, und das werden wir. Ich möchte aber kein paper und /oder Kommentar publizieren, der nur negativ ist, aus klinischen Gründen, wir muessen eine Alternative, die funktioniert, aufzeigen. Ob man ein veröffentlichtes paper zurückziehen kann, weiss ich nicht, …, wenn Du Deinen Kommentar zurückziehst, könnten wir das Korrekturpapier gemeinsam machen, so lange der Kommentar aber ohne unsre [sic] Diskussion und gemeinsames Einverständnis weggeht natürlich nicht.“

Im Laufe des Verfahrens hat die Kommission überdies Kenntnis davon erhalten, dass Prof. Birbaumer schon am 5.11.2015, also mehr als ein Jahr vor Veröffentlichung des Artikels, durch eine Email von einem ehemaligen Mitarbeiter seiner Forschungsgruppe darüber informiert wurde, dass sich aus den Daten in statistisch korrekter Auswertung keine signifikanten Ergebnisse belegen lassen, sondern dass sich bei korrekter Auswertung eine statistische Normalverteilung ergibt. Die Information war glaubwürdig sowie durch die Abbildung von Dokumenten substantiiert.

Prof. Birbaumer machte diese Feststellungen in einer Anzahl von Diskussionen über den besten Weg, die Patientendaten zu analysieren. Zu diesem Zeitpunkt – wie oben

bereits dargestellt – gab es viele Diskussionen über die Rolle von physiologisch plausiblen Modellen und ausschließlich datengetriebenen Modellen. In seiner gesamten Karriere hat Prof. Birbaumer an seine Studenten eine sokratische Einstellung vermitteln wollen: man soll sich jeder Kritik positiv annähern und Zweifel zu jeder Zeit während der Entwicklung eines Manuskripts beachten. D.h., dass er die Kritik zu diesem Zeitpunkt sehr ernst genommen hat und mit allen Mitarbeitern und Kritikern diskutiert hat. Zu diesem Zeitpunkt war ihm eine Email-Nachricht über die Berechnung von anderen Daten nicht bekannt. Was Prof. Birbaumer als Fokus im Sinn hatte, waren physiologisch plausible Modelle, wie in der a priori Hypothese festgestellt. Er betonte auch gegenüber Wissenschaftlern und Studenten von anderen Disziplinen wie den Ingenieurwissenschaften, Informatik und Psychologie, dass jede Frage, jede Kritik und jede Interpretation die physiologische Basis der erhobenen Daten respektieren muss, in diesem Fall die NIRS- und EEG-Daten. Aussagen dieser Art können nun nicht als Hinweis der Akzeptanz irgendeines spezifischen Ergebnisses interpretiert werden, sondern sie dienten der Stimulation der Diskussion und der Überprüfung auf Seiten der Mitarbeiter und Kollegen.

Wir haben keine Kenntnis von und können keine Annahme eines Berichts dieser Art im Jahr 2015 dokumentieren. Die einzigen Personen, die diese Daten zu diesem Zeitpunkt im Detail analysierten, waren Ujwal Chaudhary und Bin Xia. Wir gaben diese Daten niemals an andere Personen zur Analyse bis sehr viel später. Wir können deshalb zu diesem Punkt keine weiteren Kommentare abgeben und müssen diese Aussage als unbewiesen zurückweisen.

Daher kann jedenfalls ab diesem Zeitpunkt im Jahr 2015 nicht mehr von einem Irrtum im Sinn der fehlenden Beherrschung einschlägiger Methoden und des Fehlens statistischer Kenntnisse ausgegangen werden (vgl. o. bei Fn. 1).

Diese Informationen legen nahe, dass eine Datenverfälschung stattfand.

Wir müssen dieser Aussage widersprechen, da wir im Jahr 2015 nichts über solche Probleme wussten. Wir waren generell vorsichtig bei der vorzeitigen Annahme oder Ablehnung von Alternativhypothesen über die Daten angesichts der sehr wichtigen klinischen und ethischen Fragen, die sich in den späteren Diskussionen ergaben.

c) Daneben hält die Kommission Folgendes fest: Im Artikel, S. 19 heißt es: „If the classification accuracies … were greater than the chance-level-threshold, a new model was generated using the relative change in O_2Hb across three training sessions to give online feedback."

In der im Mai 2019 erstellten Excel-Übersicht für Patient B („PatientB_SessionDetails") heißt es dagegen zur „Feedback-session" „V2D4b3" (15.08.2014): „The model was built using the V2D4b1 and V2D4b2." Das Modell wurde also auf Basis von nur zwei „training sessions" erstellt. Gleiches gilt für die „Feedback-session" „V2D5b3" (16.08.2019), auch hier wurde auf Basis von nur zwei „training sessions" („V2D5b1" und „V2D5b2") das Modell erstellt.

Entsprechendes gilt bei Patient F für die „feedback-sessions“ „V2D6b3“ (20.05.2014) und „V3D3b3“ (06.08.2014), während bei „V2D5b4“ und „V3D4b4“ tatsächlich drei „training sessions“ herangezogen wurden. Bei Patient G wurde nur einmal das Modell aufgrund von drei „training sessions“ entwickelt, ansonsten aufgrund von nur zweien, bei Patient W nur zweimal aufgrund von nur zwei „training sessions“ (daneben einmal aufgrund von drei und einmal aufgrund von fünf). Das Modell wurde also aufgrund von zwei, drei oder fünf „training sessions“ entwickelt. Dies steht in direktem Gegensatz zu der o.g. Aussage im Artikel, S. 19.

In einigen Fällen wurden nur 2 statt 3 Trainingssitzungen verwendet. Dies war der Fall, wenn das Modell eine gute Differenzierung schon nach 2 Sitzungen ergab, was von Vorteil für die Patienten war, die nicht für lange Zeiträume trainieren können. Manchmal waren mehr Trainingssitzungen notwendig. Wir verwendeten das beste verfügbare Klassifikationsmodell, das sich manchmal früh und manchmal später im Prozess zeigte, aber im Mittel drei Sitzungen dauerte. Wir wollten im Mittel drei Sitzungen schreiben und wir können dies im Artikel korrigieren zu einem im Mittel von drei Trainingssitzungen, wenn notwendig.

5. Teilnahme am Review-Prozess bezüglich des „Formal Comment“

Dr. Spüler hat Zweifel an der Leistungsfähigkeit der von den Autoren in ihrem Artikel vorgestellten Methode in einem „Formal Comment“ formuliert, der auf PLOS-Biology am 08.04.2019 veröffentlicht wurde (https://doi.org/10.1371/journal.pbio.2004750). Dieser Text wurde von PLOS Biology zunächst auf Grund der Stellungnahme von zwei Gutachtern im Review-Prozess im Dezember 2017 zur Überarbeitung zurückverwiesen, der überarbeitete Text dann im März 2018 zurückgewiesen (vgl. Bericht, Anlage 2B) und erst auf Widerspruch von Dr. Spüler gegen diese Entscheidung publiziert (vgl. Bericht, Anlagen 3 und 4).

Nach Einleitung des Verfahrens hat Prof. Birbaumer von sich aus mehrere Dateien an die Kommission gesandt, unter denen sich ein elfseitiger Text befindet, der im ersten Satz als „Review“ bezeichnet wird („This is a review of the Comment … by Martin Spüler ...“). Auch innerhalb des oben dargestellten Review-Prozesses hat Prof. Birbaumer seine Stellungnahmen als „review“ und „re-review“ bezeichnet. Dies könnte den Eindruck erwecken, dass er trotz Interessenkonfliktes am Review-Prozess in der Rolle eines Gutachters teilgenommen habe, was im Hinblick auch § 1 Abs. 2 Nr. 2 g) VerfahrensO (willkürliche Verzögerung der Publikation einer wissenschaftlichen Arbeit als Gutachter) relevant sein könnte.

Allerdings handelt es sich hierbei nur um eine Falschbezeichnung. Prof. Birbaumer hat nicht als (anonymer) Gutachter am Review-Prozess mitgewirkt, sondern unter Namensnennung Stellungnahmen abgegeben (vgl. Bericht, Anlage 2B, Email des Herausgebers v. 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 ...“; entsprechend in der Email v. 12.03.2018, Bericht, Anlage 3, S. 2). Daher ist er nicht im Sinn der Verfahrensordnung als „Gutachter“ tätig geworden, sondern als Autor der Veröffentlichung, auf die sich die Kritik in dem „Formal Comment“ bezieht. Ein wissenschaftliches Fehlverhalten im Sinn des § 1 Abs. 2 Nr. 2 g) VerfahrensO wurde hierdurch also nicht verwirklicht.

Sowohl der Hinweisgeber als auch Prof. Birbaumer begutachteten ihre Kommentare gegenseitig, was von PLoS Biology verlangt wurde und der Terminus Begutachtung betrifft hier nur ein ausführliches Lesen und eine Evaluation des jeweiligen Artikels, nicht einen formalen Begutachtungsprozess. Das Wort "begutachten" hat im Englischen beide Bedeutungen.

B) Empfehlungen der Kommission

Nach § 16 S. 2 kann die Kommission Empfehlungen zum weiteren Verfahren abgeben, wobei Art und Schweregrad des Fehlverhaltens sowie Rechte und Interessen Dritter einzubeziehen sind.

I. Im vorliegenden Fall wirkt sich das wissenschaftliche Fehlverhalten der Betroffenen nicht nur wissenschaftsimmanent aus. Vielmehr sind die betroffenen Patienten, deren pflegende Angehörige sowie zumindest eine Krankenkasse betroffen, welche zur Finanzierung der für die von den Betroffenen entwickelten Methode notwendigen Geräte verurteilt wurde. Indem der Vorgang öffentlich wurde, hat auch das Renommee wissenschaftlicher Forschung in der Öffentlichkeit Schaden genommen. Gesellschaftliches Vertrauen in wissenschaftliche Forschung wurde enttäuscht. Auch diesen Schaden haben die Betroffenen, die seit Ende 2015 Kenntnis von der Fehlerhaftigkeit ihrer Methode hatten (vgl. o. A) III. 4. c)), zu verantworten.

II. Unter Einbeziehung dieser Aspekte gibt die Kommission folgende Empfehlungen ab:

1. Den Betroffenen Prof. Birbaumer und Dr. Chaudhary ist aufzugeben, die Publikation zurückzuziehen.

2. Die Herausgeber der Zeitschrift „PLOS Biology" sind über den Beschluss zu informieren und unabhängig davon aufzufordern, den Artikel wegen Verstoßes gegen die eigene „Data Availability Policy" zurückzuziehen.

3. Die betroffenen Forschungsförderungsinstitutionen (u.a. DFG, Volkswagen-Stiftung, Bundesministerium für Bildung und Forschung, Baden-Württemberg-Stiftung, Eva Luise und Horst Köhler-Stiftung, vgl. Artikel, S. 1f., li. Sp.) sind über den Beschluss zu informieren.

4. Der Spitzenverband der gesetzlichen Krankenkassen (GKV-Spitzenverband) sowie der Spitzenverband privaten Krankenkassen (Verband der Privaten Krankenversicherung e.V.) sind über den Beschluss zu informieren.

5. Das Rektorat sollte prüfen, ob die im Beschluss enthaltenen Feststellungen Auswirkungen auf die – sowieso nur zeitlich befristet verliehene (vgl. Richtlinie des Rektorats v. 9.11.2011, Nr. 1) – Seniorprofessur Prof. Birbaumers haben müssen. Unabhängig davon empfiehlt die Kommission dem Rektorat, davon abzusehen, bei Prof. Birbaumer weiter das in Nr. 4 der Richtlinie genannte Verfahren anzuwenden.

6. Die Inkonsistenz des der Kommission übergebenen Datenmaterials gibt ferner Anlass, sämtliche Publikationen, an denen jeder der Betroffenen seit Erhebung dieser Daten (2014) arbeitete, durch eine externe Begutachtung untersuchen zu lassen.

7. Die wissenschaftlichen Kooperationspartner Prof. Birbaumers sollten in geeigneter Weise über die in diesem Beschluss enthaltenen Feststellungen informiert werden, um ihnen die Möglichkeit zu geben, ihre Forschungsarbeit überprüfen zu können.

8. Im Sinn der im Leitbild der Universität verankerten Verantwortlichkeit („Responsibility“) regt die Kommission an, die Angehörigen der betroffenen Patienten gegebenenfalls eine Anlaufstelle anzubieten.

III. Abschließend möchte die Kommission auf die besonderen forschungsethischen Aspekte dieses Falles hinweisen und damit die Anregung an die Universitätsleitung verbinden, in den betreffenden Forschungseinrichtungen – außerhalb der dort mit möglichen Fällen wissenschaftlichen Fehlverhaltens befassten Institutionen – eine entsprechende Diskussion anzustoßen.

Die Forschergruppe um Prof. Birbaumer behauptet aus Sicht eines Laien, dass das von ihr entwickelte System Gehirndaten von Patienten im Completely-Locked-in-Zustand selbstlernend interpretieren und in Ja/Nein-Stellungnahmen übersetzen kann. Auf diesem Wege können Informationen von Seiten der Patienten in einer verlässlichen, weil wissenschaftlich verbürgten Weise aufgenommen und mit ihnen ausgetauscht werden. Die Forschergruppe stellt mithin einen wissenschaftlich validen Weg vor, mit Patienten im Completely-Locked-in-Zustand zu kommunizieren. Die Evidenz für die behauptete Kommunikation besteht einzig in der wissenschaftlichen Validität des Verfahrens. Die Ja/Nein-Stellungnahmen der Patienten können auf keinem anderen Wege als dem des BCI festgestellt und dessen Feststellung auf keinem anderen Weg geprüft werden, wobei das Kriterium der Prüfung die wissenschaftliche Qualität des angewandten Systems ist. Damit trägt die Forschergruppe die Verantwortung, die wissenschaftliche Evidenz ihrer Forschungsergebnisse zu gewährleisten und ihre Forschungsergebnisse mit allen dafür notwendigen Informationen über das Forschungsverfahren und aus dem Forschungsprozess zu belegen.

Diese Verantwortung besteht daher eben nicht nur im Rahmen der wissenschaftlichen Öffentlichkeit, sondern in besonderer Weise auch gegenüber den an ALS erkrankten Patienten, die auf den Completely-Locked-in-Zustand zugehen und sich auf das Versprechen verlassen, dass sie in diesem Zustand mit Hilfe des wissenschaftlich qualifizierten Systems kommunizieren können, sowie gegenüber denjenigen Menschen, die Patienten in diesem Zustand pflegen. Patienten und die ihnen nahestehenden Menschen können das ihnen gegebene Versprechen, miteinander kommunizieren zu können, nicht selbst überprüfen. Sie müssen einzig auf den Ausweis der Wissenschaftlichkeit des Versprechens vertrauen.

Mit der »Größe« sowie der medialen Verbreitung des Versprechens hat die Forschergruppe auch die Verantwortung für deren wissenschaftliche Bestätigung erhöht und damit auch die Verantwortung dafür, die Voraussetzungen sicherzustellen, dass ihre veröffentlichten Ergebnisse geprüft und auf diesem Wege die Behauptung der Kommunikation mit Patienten im Completely-Locked-in-Zustand verifiziert oder falsifiziert wird.

Tübingen, den 30.05.2019

Xxx (Leiter der Kommission)