MARSILIUS KOLLEG

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

JÖRG JÄCKEL
CHRISTOPH VANBERG

FEHLER MACHEN IST MENSCHLICH.

Wie gehen experimentelle Wissenschaften damit um?

https://doi.org/10.11588/fmk.2025.25.109297

MARSILIUS-KOLLEG
2023 / 2024

Christoph Vanberg

FEHLER MACHEN IST MENSCHLICH. WIE GEHEN EXPERIMENTELLE WISSENSCHAFTEN DAMIT UM?

Jörg Jäckel · Theoretische Physik
Christoph Vanberg · Economics

EINLEITUNG

Experimentelle Methoden spielen in den Natur- und auch zunehmend in den Sozialwissenschaften eine wichtige Rolle. Durch die aktive und gezielte Veränderung experimenteller Bedingungen können kausale Zusammenhänge und Wirkungsmechanismen in der physikalischen und sozialen Welt besser erforscht werden, als dies durch die passive Beobachtung natürlich auftretender Phänomene alleine möglich wäre. Damit sie diesen Anspruch erfüllen können, müssen Experimente jedoch sorgfältig entworfen und ihre Ergebnisse angemessen analysiert und interpretiert werden.

Wie in allen menschlichen Unterfangen ist es unvermeidbar, dass hierbei auch Fehler gemacht werden. Dazu gehören relativ banale Missgeschicke, wie wenn ein Reagenzglas falsch beschriftet oder einem Probanden das falsche Medikament gegeben wird. Es gehören aber auch „subtilere" Probleme dazu, wie etwa wenn bei einer Berechnung eine mathematische Methode angewendet wird, die normalerweise eine gute Annäherung erlaubt, aber im spezifischen Kontext nicht angebracht ist.

Ziel unseres gemeinsamen Projektes war es, Schwachstellen in der experimentellen Forschungspraxis sowie deren theoretischer Konzeption zu untersuchen, die Fehler begünstigen. Insbesondere interessierten uns konzeptionelle Fehler, die beim Entwurf und der Durchführung von Experimenten, sowie der Analyse der gewonnenen Daten und ihrer Interpretation auftreten können. Die interdisziplinäre Zusammen-

arbeit zwischen einem Physiker und einem Ökonomen ermöglichte es uns, gemeinsame sowie fachspezifische Fehlerquellen zu identifizieren und uns über Strategien der Fehlererkennung und -vermeidung auszutauschen.

VERLAUF UND STAND DES GEMEINSCHAFTSPROJEKTS

Zu Beginn des Projektes berichteten wir uns gegenseitig von Beispielen, die wir aus unseren jeweils eigenen Forschungsbereichen kennen. Wir waren überrascht, dass uns viele der Probleme, die im jeweils anderen Fach auftreten, bekannt vorkamen. So versuchten wir, eine systematische Klassifizierung von Fehlerquellen zu erstellen, die in beiden Fächern auftreten.

Aus unserer Sicht finden sich dabei (mindestens) folgende Kategorien: (1) „Falsch gerechnet": Die Herleitung beobachtbarer Phänomene aus der zu prüfenden Theorie enthält Fehlschlüsse, (2) „Falsch gemessen": Im Experiment wird nicht wirklich das Phänomen beobachtet, auf welches sich die Hypothese bezieht, und (3) „Falsch interpretiert": Die erhobenen Daten werden implizit im Lichte gewisser Grundüberzeugungen interpretiert, die nicht hinterfragt werden.

Im nächsten Schritt besprachen wir, welche Prozesse in unseren jeweiligen Disziplinen dazu beitragen, derartige Fehler zu vermeiden bzw. zu erkennen. Hierzu gehört z.B. der Austausch mit fachkundigen Kollegen in der Planungsphase neuer Projekte. In beiden Disziplinen werden Forschungspapiere als Preprints öffentlich zugänglich gemacht, bevor sie einem formalen Begutachtungsprozess unterzogen werden. Dies erhöht die Chance, dass Probleme früh erkannt werden - birgt aber auch die Gefahr, dass irreführende Ergebnisse kursieren. Ein interessanter Unterschied zwischen den Disziplinen ist, dass in der Physik schon der Entwurf eines neuen Experiments üblicherweise als eigenständiges Papier erscheint und vor dessen Umsetzung diskutiert wird. Diese Praxis ist in der Ökonomik (noch) unüblich. Ein weiterer Unterschied betrifft die Rolle der formalen Begutachtung durch wissenschaftliche Zeitschriften. Innerhalb der Ökonomik wird die Veröffentlichung einer Studie in einer renommierten Zeitschrift als starker Hinweis auf dessen Glaubwürdigkeit verstanden. In der Physik hingegen spielen eher Diskussionen in der Community eine wesentliche Rolle.

Eine besondere Herausforderung erscheint uns der Umgang mit bereits publizierten Studien, deren Ergebnisse aufgrund möglicher Fehler angezweifelt werden. Ein

wesentlicher Punkt ist dabei, dass es oft einen erheblichen Aufwand bedeutet, einen vermuteten Fehler zweifelsfrei nachzuweisen und dann mögliche Korrekturvorschläge zu machen. Dieser Aufwand wird leider nicht immer honoriert. Daher kommt die akribische Kritik bestehender Forschungsergebnisse teilweise zu kurz.

Zu Beginn des ersten Fellowship Semesters hatten wir die Gelegenheit, in einem jeweils eigenen Vortrag die ersten Ansätze dieser Überlegungen mit den anderen Fellows zu besprechen. Der interdisziplinäre Austausch im erweiterten Kreis bot neue Impulse und warf grundsätzliche Fragen über die erkenntnistheoretischen Implikationen der Problematik auf. Ist eine Prüfung wissenschaftlicher Theorien überhaupt möglich, wenn Fehler nicht verlässlich ausgeschlossen werden können? Kann jede Theorie vor Falsifikation geschützt werden, indem auf widersprechende Evidenz mit einem Verweis auf mögliche Fehler reagiert wird?

Es zeigt sich: Die Möglichkeit der Falsifikation einer theoretisch relevanten Hypothese erfordert die Akzeptanz mehrerer „Hilfshypothesen", unter anderem, dass keine Fehler in der Konzeption, Durchführung und Datenauswertung aufgetreten sind. Diese Tatsache erscheint uns ein Spezialfall des in der Wissenschaftstheorie bekannten *Duheme-Quine* Problems zu sein. Überspitzt formuliert ist die Aussage, dass eine eindeutige Falsifikation unmöglich ist – ein schwerer Schlag für ein von Popperschem Falsifikationismus geprägtes Wissenschaftsverständnis. Was folgt daraus für die Möglichkeit wissenschaftlichen Fortschritts? Legen wir die Hände in den Schoß und sagen, experimentelle Forschung habe sowieso keinen Sinn? Offensichtlich nicht.

Stattdessen müssen wir uns fragen, wie wir mit dem Problem *pragmatisch* umgehen. Dazu gehören aus unserer Sicht folgende Aspekte. (1) In Wissenschaft und Öffentlichkeit sollte ein Bewusstsein für die Unvermeidbarkeit von Fehlern gefördert werden, um eine naive Interpretation empirischer Forschung zu vermeiden. Es sollte darauf hingewiesen werden, dass empirische *Resultate* unvermeidbar auf zusätzlichen Annahmen beruhen und insofern immer eine *Interpretation* von Beobachtungen darstellen. (2) Bei der Bewertung empirischer Forschungsarbeit sollten Leser stets ein Auge auf die „Hilfshypothesen" werfen, auch wenn diese nicht im Vordergrund des Interesses liegen. Hierzu gehört die akribische Prüfung auf konzeptionelle und andere Fehler. (3) Wenn ein Experiment dafür kritisiert wird, dass eine Hilfshypothese nicht erfüllt ist, sollte diese Kritik anders behandelt werden, als wenn die erklärende

Theorie selbst angezweifelt wird. Insbesondere sollte nicht verlangt werden, dass eine alternative Erklärung für die Beobachtungen im fehlerhaften Experiment angeboten wird. (4) Mechanismen, die aus der Interpretation von experimentellen Ergebnissen abgeleitet werden, sollten durch unabhängige, idealerweise konzeptionell andersartige Tests überprüft werden.

Ein pragmatischer Umgang mit der potenziellen Fehlerhaftigkeit empirischer Forschung erfordert daher vor allem, dass die wissenschaftliche Gemeinschaft dieses Problem wahrnimmt. Um zu prüfen, inwiefern dies gegeben ist, haben wir für beide Teilprojekte jeweils eine Umfrage unter Kollegen durchgeführt, die wir in den folgenden Abschnitten vorstellen.

TEILPROJEKT VANBERG: SUBJEKTIVE WAHRNEHMUNG UND INTERPRETATION VON VERHALTEN IM KONTEXT ÖKONOMISCHER EXPERIMENTE

In der Diskussion mit meinem Projektpartner sowie mit den übrigen Fellows kam ich immer wieder zu einem Thema zurück, das ich in meinem zweiten Vortrag im Fellow Seminar thematisieren durfte. Dies betrifft die mögliche Diskrepanz zwischen der experimentellen *Situation*, welche ein Wissenschaftler herzustellen beabsichtigt, und der Situation, welche die Teilnehmer *wahrnehmen*. Das beobachtete Verhalten wird als Reaktion auf ein bestimmtes Problem interpretiert. Aber was ist, wenn die Probanden die Situation ganz anders verstanden haben?

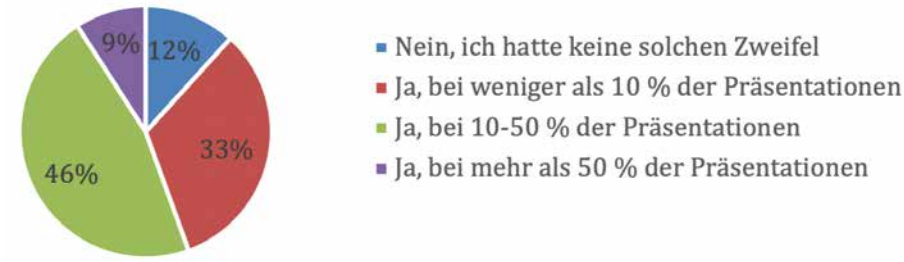
Dieses Problem ist eine Besonderheit experimenteller Forschung zu menschlichem Verhalten. Anders als bei physikalischen Teilchen hängt das Verhalten von Menschen nicht direkt von *objektiven* experimentellen Bedingungen ab, sondern von der *subjektiven Wahrnehmung* der Probanden. Was die gemeinsame Kategorisierung der Fehlerquellen betrifft, fällt dieses Problem sowohl in die Kategorie „Falsch gemessen" als auch „Falsch interpretiert": Das gemessene Verhalten ist eine Reaktion auf ein anderes Problem als das beabsichtigte, und die Ergebnisse werden durch eine *falsche Brille* betrachtet, nämlich der impliziten Annahme, dass die Situation einem bestimmten theoretischen Rahmen entspricht.

Im Austausch mit den Marsilius-Fellows gewann ich den Eindruck, dass dieses spezifische Problem als wichtig und interessant betrachtet wurde. So schien es mir

ein natürlicher nächster Schritt zu sein, das Bewusstsein für diese Problematik unter Fachkollegen zu eruieren. Zu diesem Zweck führte ich eine Umfrage unter den Mitgliedern einer internationalen Gemeinschaft experimenteller Ökonomen durch.

Die Umfrage wurde über eine Mailingliste der Gesellschaft beworben und umfasste drei Fragen. Mit der ersten Frage wollte ich das Problembewusstsein in Bezug auf noch nicht publizierte bzw. laufende Forschungsprojekte messen. Die zweite Frage untersuchte, inwiefern aktiv überprüft wird, ob die angesprochene Diskrepanz vorliegt. Die letzte Frage bezog sich auf publizierte Studien und deren Verlässlichkeit. Insgesamt haben 153 Forschende an der Studie teilgenommen. Im Folgenden fasse ich die Ergebnisse kurz zusammen. (Die Fragen und Antworten sind jeweils aus dem Englischen übersetzt.)

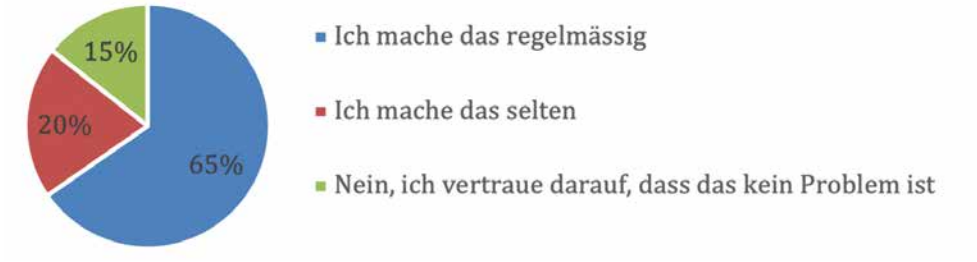
Frage 1: "Denken Sie an *Präsentationen* experimenteller Forschung, die Sie im letzten Jahr gesehen haben, z. B. bei der [Jahreskonferenz der Gemeinschaft]. Hatten Sie manchmal Zweifel an der Übereinstimmung zwischen der Problemsituation, die im Vortrag beschrieben wurde, und der Situation, welche die Versuchspersonen wahrgenommen haben?"



Bemerkenswert ist, dass nur 12% der Teilnehmer angaben, niemals derartige Zweifel zu haben. Mehr als die Hälfte der Teilnehmer ist der Meinung, dass diese bei mehr als 10% der Präsentationen angebracht sind. Dieses Ergebnis deutet auf ein ausgeprägtes Problembewusstsein in Bezug auf noch nicht publizierte Forschungsprojekte hin.

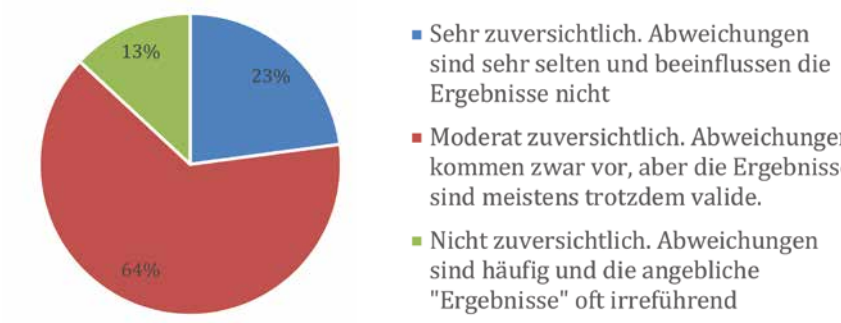
Frage 2: "Denken Sie daran, wie Sie üblicherweise ein experimentelles Forschungspapier lesen. Schauen Sie dabei in der Regel auf die Instruktionen? Prüfen Sie explizit, ob die Entscheidungssituation, wie sie von den Versuchsper-

sonen wahrgenommen wird, der im Hauptteil der Arbeit beschriebenen Situation entspricht?"



Nur 15% der Teilnehmer gaben an, dass sie auf die Übereinstimmung zwischen dem beschriebenen und wahrgenommenen Problem vertrauen. Fast 2/3 der Teilnehmer gaben an, regelmäßig auf diese Problematik zu achten. Auch dieses Ergebnis suggeriert ein ausgeprägtes Problembewusstsein und einen entsprechend kritischen Umgang mit experimentellen Studien.

Frage 3: "Denken Sie nun an experimentelle Artikel, die in renommierten Fachzeitschriften veröffentlicht wurden und häufig zitiert werden. Wie zuversichtlich sind Sie, dass die von den Versuchspersonen wahrgenommenen Probleme im Allgemeinen mit denjenigen übereinstimmen, die von den Autoren im Hauptteil der Arbeit beschrieben werden?"



Die Antworten auf die letzte Frage legen nahe, dass die meisten der Befragten größeres Vertrauen in Forschung haben, die erfolgreich publiziert und zitiert wird. Nur 13% der Befragten sind der Meinung, dass irreführende Resultate häufig vorkommen. Allerdings ist eine deutliche Mehrheit der Meinung, dass Abweichungen nicht selten sind.

Insgesamt kann man aus diesen Ergebnissen erkennen, dass die angesprochene Problematik von experimentellen Ökonomen wahrgenommen wird. Es lohnt sich, über den Umgang mit diesem Problem weiter nachzudenken.

TEILPROJEKT JÄCKEL: FEHLER IN EXPERIMENTEN UND VORSCHLÄGEN ZUR SUCHE NACH NEUEN LEICHTEN TEILCHEN

In der Teilchenphysik geht es darum, grundlegende Gesetzmäßigkeiten der Natur zu ergründen. Diese sollten objektiv und universell sein und gefundene Ergebnisse sollten Bestand haben (selbstverständlich können sie später präzisiert werden). Dementsprechend ist es besonders kritisch, wenn sich Fehler einschleichen und gegebenenfalls sogar lange Zeit unerkannt bleiben.

In den Diskussionen mit meinem Projektpartner hat sich ergeben, dass wir gerne das Problembewusstsein, aber auch das Ausmaß solcher Probleme genauer ergründen würden. Deswegen habe auch ich eine kleine Umfrage unter Kollegen innerhalb des Felds gemacht. Dazu habe ich mir bekannte Kollegen aus meinem Forschungsbereich direkt per E-Mail kontaktiert und ihnen den Link der Umfrage zugeschickt.

Diese Umfrage beinhaltete Fragen zur Bekanntheit und Häufigkeit solcher Fehler (sinngemäß ggf. auszugsweise aus dem Englischen übersetzt) wie "Kennen Sie Artikel die ein Experiment vorschlagen, die konzeptionelle oder rechnerische Probleme aufweisen?" (50% ja) und "Welcher Anteil der Papiere (publiziert oder unpubliziert) hat solche Probleme?" (mehr als die Hälfte denkt, es ist mehr als 1%). Obwohl mit deutlich geringerer Statistik, zeigen auch hier die Antworten, dass es ein klares Problembewusstsein innerhalb der Community gibt.

Eine weitere Frage nach Beispielen förderte einige Verdachtskandidaten zu Tage, zeigte aber auch (in Kombination mit persönlichen Gesprächen mit Kollegen), dass es nicht immer einfach und insbesondere zeitlich machbar ist, den Fehler explizit aufzuzeigen. Dies gilt insbesondere für Fehler, die über einfache Rechenfehler hinausgehen.

Aus den Fragen zur Detektion solcher Fehler: "Denken Sie, der Referee Prozess ist gut darin, solche Fehler zu detektieren?" (mehr als 70% nein) und "Welche Art von Fehlern sind häufiger und welche davon können durch das Refereeing besser detek-

tiert werden?" ergaben sich Hinweise, dass der Referee Prozess in seiner derzeitigen Form nicht immer alle Fehler aufdeckt und dass dies eher bei "größeren Fehlern" der Fall ist. Dennoch waren die Teilnehmer optimistisch, dass die "Diskussionen bei Workshops und Konferenzen" solche Fehler aufdecken können (etwa zwei Drittel).

Zusammengefasst gibt es klare Ähnlichkeiten in den Ergebnissen zu der Umfrage meines Projektpartners in Bezug auf das Problembewusstsein. Interessanterweise ist aber anscheinend das Vertrauen in den Refereeprozess als Qualitätskontrolle von Publikationen in der Physik deutlich geringer (zumindest im Bereich der Suche nach neuen leichten Teilchen). Sowohl die Gemeinsamkeiten als auch die Unterschiede im Umgang der Community bieten Ansätze für weitere spannende Diskussionen und Gespräche mit meinem Projektpartner.

Im Laufe des Projekts bin ich außerdem auf ein Beispiel aufmerksam geworden, dass zwar in Teilen physikspezifisch ist, aber trotzdem interessante Aspekte im Umgang mit möglichen Fehlern aufzeigt. In Kürze: In der Theorie der starken Wechselwirkung kann ein sogenannter *theta-Term* auftreten. In den Standardrechnungen verursacht dieser ein elektrisches Dipolmoment des Neutrons (naiv eine kleine räumliche Trennung von positiven und negativen Ladungen). Dieses wird aber in hochpräzisen Messungen nicht beobachtet, so dass der zugehörige Parameter mit einer Genauigkeit von mehr als einem in 1 Milliarde Teilen null sein muss. Dies motiviert das sogenannte Axion, das auch als Kandidat für die dunkle Materie fungiert. Dementsprechend gibt es eine ganze Reihe von Experimenten, die nach diesem Teilchen sucht. In den letzten Jahren hat eine Gruppe von Autoren die Berechnung des Dipolmoments aus dem theta-Term in Zweifel gezogen[1] und vertritt die Ansicht, dass dieser kein Dipolmoment verursacht und dementsprechend kein Axion benötigt wird. In Anbetracht der Tatsache, dass das Axion die Grundlage für eine ganze Reihe von Experimenten ist, ist dies klarerweise ein durchaus relevantes Problem.

In einer ganzen Reihe von Diskussionen mit Kollegen aus diesem Bereich hat sich eine klare Mehrheit überzeugt gezeigt, dass dieses Ergebnis falsch ist und in der Tat ein Axion benötigt wird. Allerdings gibt es meines Wissens noch keine komplett stringente Argumentation, für die die Autoren des Artikels keine Gegenargumente haben. Interessant ist dabei, dass hier eine konträre Mischung aus hochgradiger Relevanz, schwieriger Mathematik und Problemstellung aber auch fehlender Zeit und Anreiz (da das Ergebnis für falsch und damit nicht für sehr dringend gehalten

wird) vorliegt. Obgleich meine eigene Untersuchung der Fragestellung innerhalb der Marsilius-Zeit auch noch zu keinem endgültigen Erfolg geführt hat, plane ich dennoch mich weiter mit der Frage zu beschäftigen. Darüber hinaus stellt sich aber auch die allgemeinere Frage, wie solche Probleme in Zukunft innerhalb der Community schneller (immerhin besteht das Problem schon seit mehr als vier Jahren) angegangen werden können. Diese Frage nach dem pragmatischen Umgang mit dem Problem möchte ich auch weiterhin mit meinem Projektpartner untersuchen.

KURZES RESÜMEE UND AUSBLICK

Experimentelle Forschung und auch deren theoretische Vorbereitung ist ein menschliches Unterfangen, daher lassen sich Fehler nicht ausschließen. Dies stellt eine ernsthafte Herausforderung für den Fortschritt der empirischen Natur- und Sozialwissenschaften dar. Ein pragmatischer Umgang mit diesem Problem erfordert zunächst ein entsprechendes Bewusstsein sowie geeignete Prozesse, um Fehler frühzeitig zu vermeiden, sie zu erkennen und entsprechende Kritik zu honorieren.

Die Marsilius-Fellowship bot uns die spannende Möglichkeit, diese Problematik interdisziplinär zu diskutieren. Sowohl in der bilateralen Zusammenarbeit als auch in der Diskussion mit den übrigen Fellows haben wir viel über Unterschiede und Gemeinsamkeiten in Fehlerquellen und Strategien zum Umgang gelernt. Diesen Austausch, sowie auch die Auseinandersetzung mit den Projekten der anderen Fellows, haben wir als äußerst bereichernd empfunden.

In unseren jeweiligen Teilprojekten konnten wir bereits ein recht klares Bild bezüglich des Problembewusstseins in den jeweiligen Fachkreisen zeichnen. Auch wenn noch keine rigorosen statistischen Ergebnisse vorliegen, hat unser gemeinsames Projekt gezeigt, dass es, entsprechend der menschlichen Komponente, in beiden Wissenschaftsbereichen durchaus immer wieder Fehler in Experimenten oder Experimentvorschlägen gibt. Diese Fehler liegen nicht nur in der statistischen Analyse oder statistischen Schwankungen, sondern können durchaus sowohl rechnerischer als auch konzeptioneller Natur sein. Durch Umfragen hat sich in beiden Bereichen gezeigt, dass dies der Community durchaus bewusst ist. Zwar gibt es ein großes Maß an Selbstkorrektur durch Diskussionen, Refereeprozesse, Workshops und Konferenzen, aber gleichzeitig erscheint dies nicht immer in systematischer Weise zu geschehen. Auch darf der oft hohe Aufwand (und die relativ geringe wissen-

schaftliche Honorierung dieser Leistung) zum einwandfreien Nachweis und zur Korrektur von möglichen Fehlern nicht unterschätzt werden.

Damit ergibt sich auch eine interessante Frage für den weiteren gemeinsamen Austausch: Wie können wir als wissenschaftliche Community konzeptionelle Fehler noch besser erkennen und möglichst effizient und frühzeitig beheben? Dieser Frage wollen wir jeweils in einer weiteren Umfrage nachgehen, in der wir Fachkollegen explizit nach Lösungsansätzen fragen und dabei auch auf die jeweils andere Fachkultur hinweisen. Auch wenn das Fellowship-Jahr vorbei ist, geht die Kollaboration weiter. Wir freuen uns darauf und bedanken uns beim Kolleg und den anderen Fellows für diese einzigartige Gelegenheit zum interdisziplinären Austausch.

[1] W. Y. Ai, J. S. Cruz, B. Garbrecht and C. Tamarit, *"Consequences of the order of the limit of infinite spacetime volume and the sum over topological sectors for CP violation in the strong interactions"*, Phys. Lett. B 822 (2021), 136616, [arXiv:2001.07152 [hep-th]]. Sowie weitere Artikel aus derselben Gruppe von Autoren.

Jörg Jäckel