(16) Was heißt „wissenschaftlich bewiesen“? – Das Replikationsproblem in der Forschung

Oder: Warum wir glauben, dass soziale Vorbilder die Stimmung beeinflussen („social priming“) aber nicht, dass Homöopathie wirkt oder Telepathie funktioniert

Wenn wir sagen, etwas sei „wissenschaftlich bewiesen“, dann meinen wir meistens, dass eine Serie von Bedingungen erfüllt ist, und zwar mindestens folgende:

1. **Ein Phänomen muss mit Methoden, die derzeitigen wissenschaftlichen Standards genügen, dokumentiert sein.** Am besten ist es, wenn dies eine experimentelle Methode ist, also eine, bei der der Forscher einer Kontrollbedingung erzeugt hat und gezeigt hat, dass der interessierende Effekt in einer Experimentalbedingung auftritt, nicht aber in der Kontrollbedingung. Das tun etwa tierexperimentelle, klinische oder psychologisch experimentelle Untersuchungen. Dann wird Ratten oder Mäusen, oder Menschen, etwas gegeben, was Kontrollen nicht erhalten, und eine Variable gemessen, an der der Effekt ablesbar ist.
2. **Die Untersuchung muss publiziert und damit für alle verfügbar und prüfbar sein.** Dieser Begriff ist gerade heute, zu Zeiten wo jeder alles rasch und kostengünstig im Internet veröffentlichen kann, nicht ganz leicht zu definieren. In der Regel ist die operationale Definition von „publiziert“: ein wissenschaftliches Ergebnis wurde einer begutachteten („peer reviewten“) Fachzeitschrift vorgelegt. Die Gutachter der Zeitschrift haben das Manuskript geprüft und gefunden, dass Stil, Inhalt und Methode momentan akzeptierten Standards und Gepflogenheiten entspricht; dass Methode und Ergebnisse geeignet sind, die Schlussfolgerungen der Autoren zu stützen, und die Herausgeber der Zeitschrift waren der Meinung, das Ergebnis sei interessant genug für die

Leser der Zeitschrift. Außerdem muss die Zeitschrift von ihren Standards her von bibliographischen Fachleuten so bewertet werden, dass sie und damit die dort veröffentlichten Aufsätze in einschlägigen Datenbanken geführt werden.

Damit können Menschen, die nach wissenschaftlichen Ergebnissen suchen, diese auch finden. Ein irgendwo veröffentlichter Befund, der etwa in einer Zeitschrift auftaucht, wo keine Fachleute den Inhalt prüfen, gilt nicht unbedingt als „wissenschaftlich publiziert", und die reine Verfügbarkeit im Internet ist ebenfalls kein Kriterium für „publiziert" in einem wissenschaftlichen Sinne. Der Umkehrschluss gilt allerdings auch: Nur weil etwas in einem wissenschaftlichen Organ publiziert wurde, ist es nicht notwendigerweise schon richtig, akzeptabel oder gar bewiesen. Peer Review, das haben viele Untersuchungen gezeigt, kann sich täuschen. Es ist eigentlich nichts anderes als ein sozialer Filter und kann allenfalls sagen, was derzeit akzeptabel, vermittelbar und für andere interessant und verständlich ist. Nicht mehr und nicht weniger. [1] Dabei gilt es besonders zu beachten, dass seit ein paar Jahren ein ganzer Wald von vermeintlich peer-reviewten Online-Journals aus dem Boden geschossen ist, die nichts anderes als Gelddruckmaschinen für die Herausgeber sind. Denn die Autoren zahlen zum Teil nicht unerhebliche Summen für die Publikation. Der Peer-Review existiert nur auf dem Papier, und es ist leicht möglich, alles dort zu publizieren, auch offensichtlich falsche Ergebnisse, wie einer Untersuchung gezeigt hat [2]. Der amerikanische Bibliothekar Beall unterhält eine Liste solcher Herausgeber und Journals im Internet.

3. **Das behauptete Phänomen muss robust sein und damit mehr als eine zufällige Schwankung im Meer unserer kollektiven Wahrnehmung.** Wir kennen das aus unserem Alltag: wir meinen irgendwas spezielles gesehen zu haben und sagen dann zu unserer Partnerin: „Schau mal, da sitzt ein Steinbock im Gras". Weil wir unsere Brille nicht dabei haben, verwechseln wir einen Stein mit einem Steinbock. Unsere Partnerin hat bessere Augen und kann den Wahrnehmungsfehler aufklären. So ist das in der Wissenschaft auch häufig. Eine Arbeitsgruppe oder ein Forscher entdeckt was, vielleicht sogar mit akzeptierten Methoden nach neuestem Standard. Er kennt sich mit der Methode gut aus und ist sich also sicher, dass es sich nicht um einen Fehler, ein Artefakt oder eine Wahrnehmungstäuschung handelt.

> Auch seine Kollegen, die den Befund für eine wissenschaftliche Zeitschrift begutachten, sind seiner Meinung. Die Studie wird publiziert. Ist es deswegen schon „wissenschaftlich bewiesen“? Natürlich nicht. Damit ist gerade einmal der Diskurs eröffnet. Jemand sagt: „Schaut her, ich hab was interessantes“. So wie ich im Sommer zu meiner Frau gesagt habe: „Schau mal, da sitzt ein Steinbock in der Wiese“. Wie wird aber nun aus diesem ersten, vielleicht sogar einmaligen Befund eine wissenschaftliche Tatsache? Damit wollen wir uns heute beschäftigen. Denn nur durch Replikation, möglichst durch unabhängige Replikation und auch durch Replikation unter erweiterten und erschwerten Bedingungen und durch die anschließenden Kommunikationsprozesse in der wissenschaftlichen Literatur und der wissenschaftlichen Gemeinschaft wird aus einem Befund eine wissenschaftliche Tatsache. Die soziale Seite sparen wir uns für einen anderen Themenblog. Diesmal geht es vor allem um die Replikation.

„Replikation“ bedeutet so etwas wie: ein anderer sieht mit seinen Augen auf das gleiche Phänomen. Wenn auch er oder sie einen Steinbock sieht und noch mehr andere Leute auch, dann sitzt dort höchstwahrscheinlich wirklich ein Steinbock. Wenn ich alleine glaube einen Steinbock zu sehen, und alle anderen sagen: das ist doch ein normaler Stein, oder ein Baumstumpf, dann liege ich wahrscheinlich falsch und muss zum Augenarzt. Der Begründer der Gestalttherapie Fritz Perls pflegte zu sagen: „Wenn einer sagt ‚Du bist ein Affe‘, kannst Du es ignorieren. Wenn es zwei oder drei sagen, dann wird es Zeit, dass Du Dir eine Packung Erdnüsse kaufst.“ [3] Das ist also Replikation: das wiederholte Feststellen eines Sachverhalts, idealerweise mit ähnlichen Methoden, aber aus unterschiedlicher Perspektive.

In der Wissenschaft bedeutet Replikation je nach Fachgebiet etwas sehr anderes. In der Physik zum Beispiel, die ihre diesbezügliche Krise schon lange hinter sich hat, werden experimentelle Ergebnisse erst ernstgenommen, wenn sie multipel repliziert und geprüft, von mehreren Arbeitsgruppen bestätigt und in einem kollaborativen Prozess abgesichert sind. Erst wenn etwas „5 Sigma“ hat, also 5 mal über den Standardfehler einer Messung hinausgeht, meist durch multiple Messungen belegt ist, beginnt man etwas als tatsächlich vorhanden anzuerkennen. Das ist leider in anderen Bereichen, in der Medizin und der Psychologie, nicht so. Eben erst hat sich Richard Horton, der Herausgeber der

medizinischen Fachzeitschrift Lancet, bitter beschwert, dass in der Medizin viel zu viel Zufallsbefunde als „wissenschaftlich“ gehandelt werden: „What is medicine’s 5 sigma?“, hat er gefragt und dabei die Ergebnisse eines internationalen Workshops wiedergegeben, der sich der Replikationsproblematik angenommen hat [4].

Damit hat Horton ein Thema aufgegriffen, dass schon seit einer Weile in der Medizin schwelt: vor einigen Jahren hatte der griechische Epidemiologe Ioannidis, der in Stanford lehrt, Aufsehen erregt mit einem Aufsatz „Why most published research findings are false“ [5]. Der Artikel wurde 1,4 millionenmal angesehen und bislang 1728mal zitiert (zum Vergleich: der am meisten zitierte Artikel in der gleichen Zeitschrift im gleichen Jahrgang wurde 436mal zitiert). Ich halte diesen Aufsatz für einen der bedeutendsten methodischen Beiträge, der mir in langer Zeit untergekommen ist. Darin baut Ioannidis ein einfaches Argument auf: Autoren wollen vor allem wegen positiver Entdeckungen in Erinnerung bleiben. Forschungsergebnisse, die nicht den Erwartungen entsprechen, werden nicht zur Veröffentlichung vorbereitet, oder werden von den Herausgebern von Zeitschriften abgelehnt. Sehr häufig gibt es auch zu wenige ganz unabhängige Replikationen und die die vorliegen, werden vom selben Team oder von Leuten gemacht, die systematische Fehler, die vorher gemacht wurden, wiederholen.

Zusammen mit der Tendenz, negative Ergebnisse unpubliziert zu lassen, die vor allem am Beispiel verschiedener von der Industrie gesponserter Studien mittlerweile belegt ist – siehe Teil 8 und Teil 14 meiner Methodenserie – ergibt sich daraus eben eine explosive Mischung: erste, positive Befunde werden groß hinausposaunt, die Presse hilft dabei mächtig mit, denn auch sie ist an Neuigkeiten und Erfolgsgeschichten interessiert (siehe Teil 7 der Serie). Negative Befunde haben es anschließend sehr schwer, ernst genommen oder publiziert zu werden und wenn sie publiziert werden, erscheinen sie meistens in zweit- und dritt-rangigen Zeitschriften, die weniger oft gelesen werden. Oft werden sie gar aktiv zurück gehalten. Daher haben die meisten Leute, auch die Fachleute, oft nur die gut bekannten ersten, positiven Ergebnisse im Kopf, den Rest ignorieren sie.

Das liegt unter anderen auch daran, dass wir alle Bayesianer sind (siehe Serie Teil 5): Solche starken ersten Befunde verändern unsere Vormeinung, die anschließend wie ein Filter fungiert. Daher sehen wir vor allem das, was wir kennen und erwarten, den Rest ignorieren wir. Das ist im normalen Leben genauso wie in der Wissenschaft. Aus diesem Grunde hat die Cochrane

Collaboration, ein Netzwerk interessierter Wissenschaftler die systematisch die wissenschaftliche Erkenntnis im klinischen Bereich zusammentragen, auch festgelegt, dass ein Review nach Möglichkeit alle, auch die unpublizierte oder schlecht publizierte sog. „graue“ Literatur, einbeziehen soll. Das sind auch Diplomarbeiten, Magisterarbeiten, interne Berichte, Promotionsarbeiten und ähnliches, also alles, was nicht von den Zitationsdatenbanken erfasst ist. Wenn man das tut, dann bleibt von der vermeintlichen Klarheit und wissenschaftlichen Beweislage oft wenig übrig. So haben etwa El Dib und Kollegen im Jahr 2007 1016 zufällig ausgewählte Cochrane-Reviews untersucht mit der Frage, was wir nun wirklich wissenschaftlich sicher wissen [6]. Wenn man davon ausgeht, dass die Wissenschaftler in der Cochrane Collaboration sich erst der dringenden Fragen annehmen, dann dürfte dieser Befund repräsentativ – und vielleicht sogar noch schmeichelhaft – für die Gesamtlage der wissenschaftlichen Erkenntnis im klinisch-medizinischen Sektor sein. Nur 3.4% all dieser Reviews kommen zu einem wirklich klaren Ergebnis, ob die untersuchte Intervention wirkt oder nicht. Bei 2% war klar, dass die Intervention schädlich ist, bei 1.4% der untersuchten Interventionen war klar, dass sie wirklich gut ist. Das sind gerade mal 14 der 1016 untersuchten Interventionen! Und das Verhältnis von nützlich und wirksam zu klar schädlich ist ungünstig. Und der Rest? Bei 48% der untersuchten Interventionen wissen wir noch immer zu wenig und man muss weiter forschen. Bei weiteren 5% ist anzunehmen, dass weitere Forschung belegt, dass die Intervention schädlich ist. Und bei 43% – immerhin – kann man annehmen, dass die Intervention vermutlich hilfreich ist, aber eben zu wenig klar belegt.

Nun ist es genau dieser große Graubereich, um den es geht. Denn hier wäre eine klar negative Replikation oder eine negativ ausgegangene Studie, die nicht publiziert wird, das Erkenntniselement, das unsere Sicht der Dinge verändern könnte. Die Chancen, dass es in dem einen oder anderen Bereich unpublizierte Ergebnisse mit „negativem“ Ergebnis gibt, sind relativ hoch, aber nicht zu bemessen. Aus genau dem Grund sind Replikationen, und zwar vor allem unabhängige Replikationen, wichtig. Was heißt hier „Replikation“ und was „unabhängig“? Mein Kollege Stefan Schmidt hat sich vor einigen Jahren mal dieser Thematik angenommen und eine Übersichtsarbeit verfasst [7]. Er hat dabei festgestellt, dass zwar alle von Replikation reden, alle sie fordern, alle denken, dass es sie gibt, dass aber die wenigsten Gebiete, zumindest in der Psychologie und in den Sozialwissenschaften, wirklich gut repliziert sind – vor allem, weil Replikationen bei Forschern unpopulär und bei Zeitschriftenherausgebern

ungeliebt sind. Ausserdem sind auch Forschungsförderer nicht erpicht darauf, Replikationen zu finanzieren; sie wollen lieber dafür bekannt werden, geholfen zu haben neue Befunde ans Licht zu bringen [8]. Replikation heißt: eine Forschergruppe nimmt ein publiziertes Experiment oder eine andere Studie, baut die Methodik und alles neu nach und versucht, die Ergebnisse einer anderen Gruppe in etwa so, wie sie berichtet wurden, wieder zu finden.

Dies kann verschiedene Formen annehmen: Im engsten Sinne kann eine Replikation ganz exakt so sein, wie berichtet. Das macht kaum wer. Denn das ist ziemlich langweilig. Schon eher werden sogenannte „konzeptuelle Replikationen" gemacht. Darunter versteht man, dass man das Grundprinzip nachbaut. Wenn also jemand mit einer bestimmten Substanz, sagen wir mit der Gabe von Wasser bei Kopfschmerz, berichtet, er habe Kopfschmerzen reduzieren können [9], dann versucht man bei einer Folgestudie, diesen Befund aufzugreifen und vielleicht das Design etwas sorgfältiger zu machen: etwa besser zu messen, länger zu beobachten, die Krankheit der Patienten besser zu charakterisieren, die Intervention besser zu kontrollieren – etwa nicht nur zu sagen „trinken sie mehr", sondern den Patienten wirklich mehr Wasser zu geben und auch zu kontrollieren, dass sie mehr trinken. Dann ist es oft so, wie in unserem Beispiel, dass die Folgestudie wesentlich kleinere oder auch andere Effekte erzeugt, als die Ausgangsstudie [10]. Man sieht dann im klinischen Fall, dass eine Ausweitung des Konzeptes nicht so gut funktioniert. Man weiß aber dann nicht unbedingt, ob das „negative" Ergebnis damit zusammenhängt, dass vielleicht in der Replikationsstudie irgendein entscheidender aber vielleicht noch nicht entdeckter Parameter geändert wurde. In unserem Beispiel könnte es beispielsweise sein, dass in der ersten Studie zufälligerweise mehr Leute waren, die von Haus aus wenig trinken und bei denen die Instruktion mehr zu trinken tatsächlich positive Wirkung entfaltet, wohingegen bei einer größeren Folgestudie dieser Ausgangswertunterschied verschwindet und daher auch der Effekt.

Schon allein aus diesem Grunde sind Replikationen wichtig, vor allem konzeptuelle, um die Robustheit von Annahmen und Konzepten zu testen. Aber, wie gesagt, Replikationen sind unpopulär. Man bekommt keinen Nobelpreis dafür, dass man bestätigt, was andere gefunden haben, und schon gar nicht dafür, dass man anderer Leute Ergebnisse widerlegt. Preise erhält man für die Neuentdeckung von Befunden. Natürlich müssen diese dann, damit sie allgemein akzeptiert werden, auch von anderen Forschern repliziert worden sein. Und erst

wenn sie wirklich oft repliziert und als robust bestätigt worden sind, werden sie allgemein anerkannt. Zumindest ist das in der Theorie so und in der Praxis häufig. Aber es sind auch schon Nobelpreise für Wirtschaft etwa vergeben worden für Modelle, die zwar genial sind, sich aber dann doch nicht bewähren, wie die letzte Krise der Finanzwirtschaft gezeigt hat.

Aber kehren wir zurück zur klinischen Forschung. Dort sind Replikationen eigentlich auch nötig, damit wir eine Intervention als „wirksam" akzeptieren. Warum ist das so? Es könnte ja schließlich sein, dass eine erste Studie einfach ein zufälliger Befund ist, eine Art Zufallsschwankung. Wenn das so wäre, dann müsste beim nächsten Versuch, den Effekt festzustellen, die statistische Schwankung in die andere Richtung ausschlagen und wir würden einen Null-Effekt oder gar einen negativen Effekt sehen. Und beim dritten Versuch vielleicht gar nichts mehr, so dass über alle Studien hinweg tatsächlich ein Null-Effekt am Ende steht. Nehmen wir an, der in der ersten Studie gefundene Effekt ist ein systematisch positiver, der belegt, dass die Intervention – Wassertrinken bei Migräne in unserem Beispiel – erfolgreich ist. Dann müsste eine Folgestudie eben wieder einen solch positiven Effekt zeigen, und eine dritte wieder. Und auch wenn mal eine negative oder nur sehr schwach positive Studie dabei wäre, über alle Studien hinweg müsste sich dann ein positiver und von Null deutlich verschiedener Effekt herauskristallisieren. Das könnte man in einer Meta-Analyse, etwa so wie sie die Cochrane Collaboration publiziert, nachweisen.

Man sieht an diesem Beispiel sofort: **sobald eine negative Studie unterschlagen wird, verzerren wir das Bild**. Daher ist es so wichtig, dass alle, aber auch wirklich alle, Befunde, vor allem und gerade die negativen, publiziert werden. Denn meistens lernen wir von den negativen Befunden mehr als von den positiven. Wie man aus den oben kurz referierten Daten der Cochrane Collaboration sieht, passiert das weit weniger, als man denkt.

Was Ioannidis für die Medizin theoretisch argumentiert [5], Horton vor kurzem nochmals wiederholt hat [4] ist im Rahmen der Medizin empirisch nicht untersucht. Aber die Psychologie hat seit kurzem definitiv ein ernstes Replikationsproblem. Denn nun ist empirisch klar: weniger als die Hälfte aller in der Psychologie publizierten experimentellen Befunde sind auch nur annähernd so replizierbar, wie in der Literatur berichtet [11]. Dem Psychologen Brian Nosek war vor einiger Zeit das Problem bewusst geworden. Er begeisterte eine sehr große Gruppe von Kollegen dafür, Daten von 100 Studien aus den letzten

Jahrgängen der wichtigsten psychologischen Zeitschriften („Psychological Science“, „Journal of Experimental Psychology“, „Journal of Personality and Social Psychology“) mit exakt den gleichen Methoden zu replizieren. Er suchte sich dafür Arbeitsgruppen aus, die in den fraglichen Gebieten ausgewiesen waren und sich mit der Methode auskannten. Diese traten dann in Kontakt mit den Erstautoren und ließen sich die Methode genau erklären und andere Details, die in der Publikation vielleicht nicht dargestellt werden konnten. Die Forscher scheuten wirklich keinen Aufwand, die Originalbefunde so getreulich wie möglich zu replizieren. Die gefundenen Effektstärken waren nur halb so groß wie die original publizierten [12]. 97% der originalen Studien berichteten signifikante Effekte. Aber nur 35% der Replikationen konnten signifikante Effekte finden. Nur 47% der originalen Effektstärken, also weniger als die Hälfte, lagen innerhalb des Konfidenzintervalls [13] der Replikationen.

Mit anderen Worten: weniger als die Hälfte der original berichteten Daten war mit den Ergebnissen der Replikation statistisch kompatibel. Nur 39% der Studien wurden von den Forschern subjektiv als erfolgreiche Replikation gewertet. Die mittlere Effektstärke der original berichteten Studien war r = 0.4, die replizierte war r = 0.197, also gerade mal halb so groß. Die originalen p-Werte korrelierten negativ mit den replizierten mit r = -.327. Das bedeutet: je größer und statistisch signifikanter die original berichteten Effekte waren, desto wahrscheinlicher war es, dass sie nicht replizierbar waren. Insgesamt wird ein deutlicher negativer Publikationsbias sichtbar. Das bedeutet: bei den originalen Untersuchungen sind negative Ergebnisse nicht publiziert worden oder die Forscher haben solange probiert, bis sie positive Befunde hatten, bzw. haben nur solche Aspekte aus einer Untersuchung dargestellt, die den ganzen Befund positiv haben erscheinen lassen.

Was heißt das? **Anscheinend ist die Tendenz auch in der Psychologie weit verbreitet, negative Befunde zu unterschlagen**. Das ist gerade bei experimentellen oder Querschnittstudien leicht. Sie sind in der Psychologie – aber auch in der Medizin, der Pharmakologie oder der biologischen Grundlagenforschung – nicht sehr kompliziert durchzuführen, wenn man sich mit einer Methode auskennt und alles zur Hand hat. Alle Psychologiestudenten müssen Stunden als Versuchspersonen sammeln, damit sie zu ihren Prüfungen zugelassen werden und dienen damit als menschliche Versuchskaninchen für ihre Dozenten, die auf diese Weise mal das eine oder andere ausprobieren können.

Dann kann man rasch ein neues kleines Experiment machen, weil man mal eine gute Idee hat. Kommt dabei nichts heraus, werden die Daten ignoriert. Hat man – vielleicht zufällig oder weil geschlampt wurde – einen positiven Befund, ruft man Hurra, öffnet den Sekt und schickt ein Manuskript an eine Zeitschrift. Eine andere Arbeitsgruppe fühlt sich inspiriert, will das Ergebnis replizieren, hat aber einen negativen Befund. Diese negative Replikation wird nur selten publiziert. Und so häuft sich in der Literatur ein Sammelsurium von anscheinend positiven wissenschaftlichen Befunden an. Wie oben dargestellt: fast alle originalen Befunde berichten von gefundenen Effekten, von Zusammenhängen oder bedeutsamen Unterschieden.

Man sieht also „wissenschaftlich bewiesen" heißt nicht: „irgendwo ist ein positives Ergebnis publiziert". Es muss auch gewährleistet sein, dass dieses Ergebnis repliziert worden ist, und zwar idealerweise von einer anderen Arbeitsgruppe mit deren Methode. Und je kontroverser der Befund, umso stabiler muss die Replikation sein. Beachten sollte man den Komparativ „kontroverser": wenn Daten in ein momentan akzeptiertes Denkmodell oder in einen Theorierahmen passen, wird man schon ein oder zwei Replikationen als ausreichendes Indiz für die Richtigkeit des originalen Befundes werten.

Da der originale Befund schon mit großer Wahrscheinlichkeit eine Selektion aus allen möglichen, auch negativen Daten darstellt, ist auch bei „allgemein akzeptierten" Befunden die Wahrscheinlichkeit hoch, dass man einem Fehler aufsitzt, wenn man sagt, etwas sei „wissenschaftlich" bewiesen. So auch Ioannidis kritischer Kommentar. Handelt es sich aber um kontroverse Gebiete, wird die Forschergemeinde wirklich sehr robuste Belege, also multipel replizierte und vor allem unabhängig replizierte Befunde, erwarten.

Und das ist auch der eigentliche Grund, warum *wir* (gemeint ist der Mainstream der Gesellschaft und der Wissenschaft) davon ausgehen, dass Antidepressiva wirken, Homöopathie aber nicht und warum wir glauben, dass es „social priming" gibt, aber nicht Telepathie. Rein objektiv betrachtet, ist der meta-analytisch festgestellte Unterschied zwischen homöopathischen Präparaten und Placebopräparaten über alle bekannten, publizierten und unpublizierten Studien hinweg statistisch von Null verschieden mit einer Odds ratio von 1.53 bzw. bei den zuverlässigen Studien bei OR = 1.98. Das heißt, ein Patient, der mit Homöopathie behandelt wird, hat eine doppelt so hohe Chance oder mindestens halb so große Chance geheilt zu werden, wie einer, der mit Placebo behandelt

wurde [14]. Ich habe das in meinem letzten Beitrag der Serie diskutiert (Teil 15).

Aber das theoretische Verständnis, was bei der Homöopathie passiert, ist wissenschaftlich nicht geklärt. „Social priming" passt in den momentanen Mainstream der Sozialpsychologie. Damit ist gemeint, dass jemand, der als sozial wichtige Person angesehen wird, durch sein Verhalten nicht nur das Verhalten anderer beeinflusst, sondern auch deren Gefühle und Kognitionen [15]. Die bekannteste Studie ließ jemanden mit müdem Schritt und sichtlich unter Anstrengung an einer Gruppe von Teilnehmern vorbeilaufen und stellte anschließend bei diesen fest, dass sie auch langsamer liefen, sich müde und depressiv gestimmt fühlten. Das passt ins Paradigma der subliminalen kognitiven Steuerung, das derzeit überall untersucht wird.

Das Problem: Der Befund stellte sich als unreplizierbar heraus, und diese Erfahrung war u.a. Anlass für Brian Noseks „Open Science Collaboration Project". Allerdings sahen diese negativen Befunde nie das Licht der Öffentlichkeit, weil Zeitschriftenherausgeber die Kompetenz der Experimentatoren anzweifelten oder schlicht und ergreifend nicht publizieren wollten. Weil der Befund mit der Mainstream Meinung kompatibel ist, daher glauben „wir" kollektiv immer noch an die Möglichkeit von Social Priming in einem sehr weiten Sinne, stehen aber der Möglichkeit der Wirkung homöopathischer Arzneien sehr skeptisch gegenüber, obwohl es wesentlich mehr und wesentlich robustere Befunde gibt.

Nur: **auch bei der Homöopathie sieht es mit der unabhängigen Replikation schlechter aus**, als die Meta-Analyse von Mathie nahelegt. Denn in der Meta-Analyse wurden viele unabhängige Daten gemeinsam verarbeitet. Die wenigen Versuche, innerhalb eines Forschungsmodells Replikationen durchzuführen waren selten erfolgreich. Das ist der Grund, weswegen ich selber schon vor Jahren zu dem Ergebnis gekommen bin, dass was auch immer hier wirkt nicht in einem klassisch-kausalen Sinne wirkt. Man muss eine Reihe von Tricks anwenden, mit denen es uns am Ende auch gelungen ist, homöopathische Effekte von denen von Placebo in experimentellen Studien zu isolieren, und zwar mehrfach [16]. Nun müssen andere die Befunde replizieren, damit sie potenziell dazu beitragen können, Homöopathie als „wissenschaftlich bewiesen" dastehen zu lassen. Bevor dies geschieht, muss aber auch noch eine plausible Theorie gefunden werden.

Ähnlich ist es mit der Telephatie und der Psychokinese. Es gibt viele positive Befunde [17]. Es gibt auch mindestens ein theoretisches Modell [18] und wir

haben kürzlich ein vielversprechendes experimentelles Paradigma repliziert, worüber wir demnächst berichten werden. Die Befunde sind statistisch mindestens so robust wie viele der Mainstream-Effekte. Aber erstens ist die Theorie nicht allgemein akzeptiert und akzeptierbar [19], und zweitens sind bislang unabhängige, zielsichere Replikationen eines experimentellen Paradigmas nicht gelungen [20].

Damit dem Publikationsbias entgegengewirkt wird, sollten Studien registriert werden, bevor sie durchgeführt werden. Das erleichtert hinterher die Suche nach gemachten, aber nicht publizierten Arbeiten. Das ist in der klinischen Forschung mittlerweile verpflichtend, und die meisten Journals akzeptieren keine Publikationen von klinischen Studien mehr, die nicht in einem gängigen Studienregister registriert worden sind. In der Parapsychologie ist diese Politik, alle Studien zu publizieren, auch die negativen, schon seit mindestens 20 Jahren implementiert [19]. Aber in der experimentellen Forschung, sowohl in der Psychologie als auch in der Medizin oder Pharmakologie gibt es dazu erst sehr flüchtige Ansätze.

Was also „wissenschaftlich belegt" ist, ist sehr schwer greifbar. Publikationen alleine reichen nicht aus. Sie sind außerdem oft tendenziell falsch positiv, und eine gehörige Portion Kritik sollten wir immer walten lassen. Aber selbst wenn positive Daten vorliegen und publiziert sind, stellt sich immer die Frage: sind sie repliziert und replizierbar? Und wenn diese Frage positiv beschieden ist, müssen wir uns immer noch fragen: Sind die Ergebnisse sozial, im Rahmen der derzeit geltenden Theorien und Denkmodelle akzeptabel. Und erst wenn alles drei gegeben ist, ist etwas „wissenschaftlich belegt". Wie wir gesehen haben, werden solche Urteile, vor allem wenn es um Mainstream-nahe Theorien und Effekte geht, oft vorschnell getroffen. Und daher kann etwas „wissenschaftlich belegt sein" und trotzdem falsch. Und anderes als „wissenschaftlich fragwürdig" gelten und trotzdem am Ende richtig sein

Quellen und Hinweise

[1] Henderson, M. (2010). End of the peer review show. British Medical Journal, 340, 738-740.

Hopewell, S., Collins, G. S., Boutron, I., Yu, L.-M., Cook, J., Shanyinde, M., et al. (2014). Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. British Medical Journal, 349, g4145.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American society for Information Science and Technology, 64, 2-17.
Ritter, J. M. (2011). Impact, orthodoxy and peer review British Journal of Clinical Pharmacology, 72, 367-368.
[2] Bohannon, J. (2013). Who's afraid of peer review? Science, 342(6154), 60-65.
Walach, H. (2015). Die Schrott-Schwemme und fünf Gründe, warum wir nicht dazugehören. Forschende Komplementärmedizin, 22, 152-154. http://www.karger.com/Article/Pdf/434665
[3] Das ist ein überliefertes Diktum, das meine Gestaltlehrer John und Judith Brown, die Fritz Perls noch selber erlebt hatten, kolportiert haben.
[4] Horton, R. (2015). Offline: What is medicine's 5 sigma? Lancet, 385, 1380. http://dx.doi.org/10.1016/S0140-6736(15)60696-1
[5] Ioannidis, J. P. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124
[6] El Dib, R. P., Atallah, A. N., & Andriolo, R. B. (2007). Mapping the Cochrane evidence for decision making in health care. Journal of Evaluation in Clinical Practice, 13, 689-692.
[7] Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the Social Sciences. Review of General Psychology, 13, 90-100.
[8] *Wir hatten einmal einen Antrag zusammen mit einem norwegischen Team gestellt, in dem es um die Replikation von Befunden aus der Placeboforschung hätte gehen sollen. Dies taten wir vor dem Hintergrund, dass unsere eigenen Ergebnisse der bekannten Datenlage widersprachen (etwa Walach, H., Schmidt, S., Bihr, Y.-M., & Wiesch, S. (2001). The effects of a caffeine placebo and experimenter expectation on blood pressure, heart rate, well-being, and cognitive performance. European Psychologist, 6, 15-25; und Walach, H., Schmidt, S., Dirhold, T., & Nosch, S. (2002). The effects of a caffeine placebo and suggestion on blood pressure, heart rate, well-being and cognitive performance. International Journal of Psychophysiology, 43, 247-260.) und Placebo-Effekte mit Placebo-Koffein wesentlich geringer ausfielen, als in der Literatur berichtet. Die Gutachter sagten allen Ernstes: das Vorhaben sei nicht förderungswürdig, da es sich lediglich um eine Replikation handele.*
[9] Spigt, M. G., Kuijper, E. C., van Schayck, C. P., Troost, J., Knipschild, P. G., Linssen, V. M., et al. (2005). Increasing the daily water intake for the prophylactic treatment of headache: a pilot trial. European Journal of Neurology, 12, 715-718.

[10] Spigt, M., Weerkamp, N., Troost, J., van Schayck, C. P., & Knottnerus, J. A. (2012). A randomized trial on the effects of regular water intake in patients with recurrent headaches. Family Practice, 29, 370-375.
[11] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.
[12] *Ich habe den Begriff „Effektstärken“ in meinem Methodenblog Teil 13 ausführlich erklärt. In Kürze: es ist ein Mass für die absolute Grösse eines Effektes über verschiedene Studien hinweg, das vergleichbar ist. Es gibt Masse für den Zusammenhang, der Korrelationskoeffizient, für den Unterschied zwischen Gruppen in kontinuierlichen Massen („d“, oder „g“) und für den Unterschied zwischen Gruppen in dichotomen Massen (z.B. Odds Ratio). Sie lassen sich alle ineinander überführen. Hier wurde der Korrelationskoeffizient verwendet, der von -1 (perfekter negativer Zusammenhang) bis +1 (perfekter positiver Zusammenhang) reicht. Normalerweise ist in der Sozialforschung ein Zusammenhang von Variablen in der Größenordnung zwischen r = .3 und .5 zu erwarten: Der Zusammenhang zwischen Intelligenz und Einkommen liegt etwa in diesem Bereich oder zwischen Schulnoten und künftigem Einkommen.*
[13] *Auch das Konfidenzintervall habe ich in früheren Kapiteln erklärt. Es handelt sich dabei um eine Abschätzung, ob ein gefundener Wert innerhalb oder ausserhalb eines statistischen Schwankungsbereiches liegt, innerhalb dessen z.B. 95% der Werte zu erwarten sind. Technisch werden Konfidenzintervalle folgendermaßen ermittelt: man errechnet den Standardfehler des Mittelwertes (SEM). Das ist die Schwankungsbreite der Mittelwertschätzung, die aufgrund der statistischen Unsicherheit der Schätzung zu erwarten ist. Er ist definiert als die Standardabweichung, die um den empirischen Mittelwert errechnet wurde, dividiert durch die Wurzel aus der Gesamtzahl der Beobachtungen. Wenn diese sehr groß ist, wird der Standardfehler klein, wenn sie klein ist, ist er groß. Damit wird also die Sicherheit der Schätzung ausgedrückt. Diesen Standardfehler des Mittelwertes kann man nun interpretieren als ein Standardmaß einer Normalverteilung, die um den empirisch gefundenen Mittelwert konstruiert ist, mit dem Standardfehler als Streuungsmaß. Dann ist der Konfidenzbereich des Mittelwertes definiert als ±1.96*SEM. Auf diese Weise kann man übrigens aus Daten in publizierten Studien auch die Standardabweichung rückrechnen, falls diese nicht angegeben ist.*
[14] Mathie, R. T., Lloyd, S. M., Legg, L. A., Clausen, J., Moss, S., Davidson, J. R., et al. (2014). Randomised placebo-controlled trials of individualised homoeopathic treatment: sytematic review and meta-analysis. Systematic Reviews, 3(142).

[15] Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. Journal of Personality and Social Psychology, 81, 1014-1027.
[16] Möllinger, H., Schneider, R., Löffel, M., & Walach, H. (2004). A double-blind, randomized, homeopathic pathogenetic trial with healthy persons: Comparing two high potencies. Forschende Komplementärmedizin und Klassische Naturheilkunde, 11, 274-280.
Walach, H., Sherr, J., Schneider, R., Shabi, R., Bond, A., & Rieberer, G. (2004). Homeopathic proving symptoms: result of a local,non-local, or placebo process? A blinded, placebo-controlled pilot study. Homeopathy, 93, 179-185.
Möllinger, H., Schneider, R., & Walach, H. (2009). Homeopathic pathogenetic trials produce symptoms different from placebo. Forschende Komplementärmedizin, 16, 105-110.
Walach, H., Möllinger, H., Sherr, J., & Schneider, R. (2008). Homeopathic pathogenetic trials produce more specific than non-specific symptoms: Results from two double-blind placebo controlled trials. Journal of Psychopharmacology, 22, 543-552.
Eine Zusammenfassung bietet Walach, H., & Teut, M. (2015). Scientific provings of ultra high dilutions in humans. Homeopathy, in print.
[17] Schmidt, S. (2014). Experimentelle Parapsychologie – Eine Einführung. Würzburg: Ergon.
[18] Walach, H., von Ludacou, W., & Römer, H. (2014). Parapsychological phenomena as examples of generalized non-local correlations – A theoretical framework. Journal of Scientific Exploration, 28, 605-631.
Lucadou, W. v. (2015). The Model of Pragmatic Information (MPI). In E. C. May & S. Marwaha (Eds.), Extrasensory Perception: Support, Skepticism, and Science: Vol. 2: Theories and the Future of the Field (pp. 221-242). Santa Barbara, Ca: Praeger
[19] *Der Philosoph Daniel Dennett sagte einmal zu Dick Bierman, einem Parapsychologieforscher: wenn sich herausstellen sollte, dass solche Effekte tatsächlich existieren, würde er Selbstmord begehen. Das ist natürlich eher spassig zu verstehen, zeigt aber, wie hoch die emotionalen Wellen schlagen.* Zitat als persönliche Mitteilung, in Bierman, D. J. (2001). On the nature of anomalous phenomena: Another reality between the world of subjective consciousness and the objective world of physics? In P. Van Loocke (Ed.), The Physical Nature of Consciousness (pp. 269–292). New York: Benjamins, p. 269.
[20] Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous

retroactive influences on cognition and affect. Journal of Personality and Social Psychology, 100, 407-425. Diese Serie von positiven Befunden liess sich nicht replizieren Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: three unsuccessful attempts to replicate Bem's ‚retroactive facilitation of recall' effect. PLoS One, 7(3), e33423. Und eine große Replikation der langjährigen Versuche des PEAR-Labs schlug ebenfalls fehl: Jahn, R. G., Dunne, B. J., Bradish, G. J., Dobyns, Y. H., Lettieri, A., Nelson, R. D., et al. (2000). Mind/machine interaction consortium: PortREG replication experiments. Journal of Scientific Exploration, 14, 499-555.

← Zurück zu Kapitel 15
Weiter zu Kapitel 17 →