Einsichten eines Wissenschaftsnarren (11)

Kann denn (Nicht-) Reproduktion Schande sein ?

Wissenschaft will neues Wissen generieren. Neues Wissen muss aber reproduzierbar sein. Dumm nur, dass Reproduzierbarkeit ein heterogener Begriff ist – und keineswegs einem einfachen Ja/Nein-Schema folgt.

Die Ergebnisse Deiner Arbeit ließen sich nicht reproduzieren! Diese Schreckensmeldung fürchtet in letzter Zeit so mancher. Reproduzierbarkeit, Replizierbarkeit, Reliabilität und Robustheit der Forschung werden von den wissenschaftlichen Akademien, den Journalen, und mittlerweile auch von den Fördergebern angemahnt. Es ist eine Bewegung für „Reproduzierbare Wissenschaft" entstanden; Förderprogramme für die Reproduktion von Forschungsarbeiten sind derzeit in Vorbereitung. In einigen Wissenschaftszweigen, allen voran der Psychologie, aber auch in Feldern wie der Krebsforschung werden Forschungsarbeiten nun auch systematisch repliziert. Oder oft eben *nicht*. Deshalb erleben wir eine „Reproduzierbarkeits-Krise".

Mit Daniel Fanelli hat nun kürzlich ein Wissenschaftler mahnend seine Stimme erhoben, den man bisher auf der Seite der Befürworter solcher Aktivitäten vermutete. „In den ehrwürdigen *Proceedings of the National Academy of Sciences* fragt er rhetorisch: *„Is science really facing a reproducibility crisis, and do we need it to?"* Ich möchte mich daher heute, vielleicht am Vorabend einer aufkeimenden Gegenbewegung, mit einigen Einwänden gegen das derzeitige Mantra von der „Reproduzierbaren Wissenschaft" auseinandersetzen.

Ist Reproduzierbarkeit von Ergebnissen wirklich das Fundament der wissenschaftlichen Methode? Oder hat nicht, wie Chris Drummond anmerkt, schon Thomas Kuhn in seinem berühmten Werk „Die Struktur wissenschaftlicher Revolutionen" festgestellt, dass der wissenschaftliche Fortschritt ganz und gar *nicht* in der „normalen", durch Aufeinanderaufbauen voranschreitenden Wissenschaft stattfindet, sondern durch periodisch wiederkehrende „Paradigmenwechsel"? Und der Paradigmenwechsel ist doch alles andere als die Reproduktion von bisher Dagewesenem!

Ein verwandtes Argument ist das von der „Trivialität" reproduzierter wissenschaftlicher Ergebnisse. Danach sind gerade die Befunde, die auf sattsam Bekanntem aufbauen, garantiert die reproduzierbarsten. Und umgekehrt: Bedeutet erfolgreiche Reproduktion denn, dass es sich um *richtige* Resultate handelt? Was, wenn Originalresultat und Reproduktion demselben systematischen Fehler aufsitzen; oder wenn beide ganz zufällig falsch-positive Befunde sind?

»Komischerweise gilt ja stets das Resultat der Replikation als das richtige.«

Und Vorsicht, es wird noch philosophischer. Schließlich bezieht sich so mancher Kritiker der Betonung von Reproduzierbarkeit als Ziel von Wissenschaft gar auf Karl Popper: Nach ihm lassen sich Hypothesen nicht beweisen, sondern nur falsifizieren. Nehmen wir das berühmte Beispiel des schwarzen Schwans, der die Hypothese „Alle Schwäne sind weiß" widerlegt. Eine Studie, welche eine vorherige Untersuchung, die an einem See nur weiße Schwäne vorfand, insofern reproduziert, dass sie an einem anderen See auch nur Artgenossen mit weißen Federn findet, hat diese zwar erfolgreich repliziert – die Hypothese wäre aber trotzdem falsch. Was sich insbesondere zeigen würde, wenn der schwarze Schwan vorbeifliegt.

Dies ist, was Jason Mitchel als „Leere der misslungenen Replikation" bezeichnet. Das Tolle an Wissenschaft ist doch schließlich die Entdeckung von Neuem, nicht die langweilige Wiederholung. Reproduzieren ist also keine Wissenschaft, lautet hier das Verdikt!

Ohne solche theoretischen Umschweife gehen dagegen jene Kritiker zur Sache, die Replikations-Experimente grundsätzlich für problematisch halten – und zwar, weil sie Zweifel an der Kompetenz der Replizierer hegen. Meist verweist man dann auf die Heerscharen von Doktoranden und Postdocs, die aufgerieben wurden, um eine bestimmte Technik im eigenen Labor zu etablieren. Natürlich würde auch dort von „echten" Experten alles replizierbar sein. Aber das Vorhandensein von implizitem Wissen, das nicht im Methodenteil von Artikeln wiedergegeben werden kann, verhindere am Ende die Wiederholbarkeit. Demnach beweise die Nicht-Wiederholbarkeit der Ergebnisse durch andere folglich nur eines: deren Unfähigkeit!

Und noch etwas sehr Ernstzunehmendes führen die Kritiker ins Feld: Durch die moralische Überhöhung der Replikation als Goldstandard werden Wissenschaftler stigmatisiert, deren Ergebnisse nicht wiederholt werden können. Ganz unabhängig von den Details und Umständen der Replikation, gilt ja irgendwie stets das Resultat der Replikation als das richtige. Bei Nicht-Replikation steht daher auch gleich der Verdacht mit im Raum, dass hier jemand nicht sauber gearbeitet, ja vielleicht sogar gegen die Regeln der guten wissenschaftlichen Praxis verstoßen hat! Denn gute Wissenschaft *muss* schließlich replizierbar sein!

Haben die Kritiker also recht? Ist es ein Fehler, Reproduzierbarkeit von Forschung aufs Schild zu heben, sie zu belohnen und gar Fördermittel dafür auszugeben? Ganz sicher nicht. Trotzdem empfiehlt der Narr, die obigen Argumente ernst zu nehmen und sich mit dem nicht ganz trivialen Thema wirklich auseinanderzusetzen.

Zunächst einmal geht es unter dem Stichwort „Reproduzierbarkeit" begrifflich häufig drunter und drüber. Reproduzierbarkeit der Methoden, der Resultate, der aus den Ergebnissen abgeleiteten Schlüsse (die *inferentielle* Reproduzierbarkeit) sowie strikte Replikation, und so weiter... – das muss man alles sehr

wohl auseinanderhalten. Meinen wir eine Wiederholung der Effektgröße, des p-Wertes oder von statistischer Signifikanz überhaupt?

Und natürlich ist Reproduzierbarkeit kontextabhängig. Da steckt nämlich tatsächlich „implizites Wissen" drin. Viel wichtiger aber noch ist die Robustheit der Ergebnisse, also ihre externe Validität. Hanno Würbel hat in diesem Zusammenhang etwa auf das Paradox des Standardisierungs-Irrtums hingewiesen (*LJ* 7-8/2018: 18-21): Der Wunsch nach mehr Reproduzierbarkeit führt häufig zum Ruf nach mehr Standardisierung. Dies aber, und darin steckt das Paradox, ist ein Holzweg – denn mit höherer Standardisierung werden Ergebnisse schlechter reproduzierbar!

Schon Ronald Fisher, der Urvater der von uns so verehrten frequentistischen Wahrscheinlichkeitstheorie, hat es 1935 so formuliert: *„Ein hoch standardisiertes Experiment liefert nur direkte Informationen in Bezug auf den engen Bereich der Bedingungen, welche durch die Standardisierung erreicht wurden. Verglichen mit bewusster Variation der Bedingungen stärkt Standardisierung daher nicht unsere Schlussfolgerungen aus den Ergebnissen, sondern schwächt sie sogar."* Gerade in der biomedizinischen Wissenschaft ist diese verbesserte externe Validität aufgrund von bewusster oder unbewusster Variation – und somit unter Verzicht auf Standardisierung! – aber sehr wichtig: Wenn sich ein Ergebnis aus einer Maus in Boston in einer genetisch identischen Maus in Berlin nicht wiederholen lässt, spricht das erstmal nicht gegen die Richtigkeit und Qualität der Befunde aus Boston. Es lässt aber sehr wohl Zweifel an deren Übertragbarkeit auf den Menschen aufkommen.

Natürlich ist das Replizieren von eigenen Befunden sowie von denjenigen anderer wertvolle Wissenschaft. Zum einen – und hier liegt das Missverständnis bei der Interpretation von Thomas Kuhn – beruhen sowohl die „normale Wissenschaft" (also das, was die meisten von uns tun) als auch die Forschung, die zu Paradigmenwechseln führt (also das, was der Zufall und geniale Wissenschaftler bewerkstelligen), entscheidend auf Ergebnissen, die wiederholbar sein müssen. Dabei führt sowohl die Reproduktion als auch eine mögliche Nicht-Reproduktion zu wissenschaftlich relevanten Ergebnissen.

»Wenn Ergebnisse nicht reproduziert werden, fängt die Wissenschaft oft erst richtig an.«

Eine kompetente Reproduktion kann eine Hypothese stärken, insbesondere wenn sie auch unter Variation von methodischen Details erfolgreich war. Wird das Design der Reproduktion so verändert, dass bewusst alternative methodische Ansätze gewählt werden – zum Beispiel statt einer *Knock-out*-Maus die Manipulation des interessierenden Gens mittels RNA-Interferenz –, spricht man von Triangulation und erhält potenziell noch robustere Resultate. Andererseits kann eine Nicht-Reproduktion über das Erkennen modifizierender Faktoren zu neuen Erkenntnissen führen.

In keinem Fall darf Nicht-Reproduktion daher zur Stigmatisierung führen. Unzählige Faktoren können diese verursachen, die wesentlichen habe ich oben den Replikations-Kritikern in den Mund gelegt.

Und hier gleich noch eine Warnung an diejenigen, die die Debatte irrelevant finden, da sie „ja schon immer ihre eigenen Ergebnisse repliziert haben". Ein Effekt, der auf einem Niveau von p=0.05 gerade eben noch signifikant war, lässt sich bei strikter Replikation (gleiches Experiment, gleiche Fallzahl, *et cetera*) nur mit fünfzigprozentiger Wahrscheinlichkeit als signifikant wiederholen – selbst wenn er ein *tatsächlich wahres* Ergebnis darstellt. Ein Würfelspiel also! *(Der Wissenschaftsnarr hatte dies ausführlich in* LJ *4-2017: 24-25 dargelegt)*.

Der Narr meint daher: Replizierbarkeit ist zwar nicht der Zweck von Wissenschaft – da geht es um neues Wissen. Aber neues Wissen muss reproduzierbar sein. Karl Popper meinte hierzu: *„Alle Ereignisse, die nicht reproduzierbar sind, sind aus der Wissenschaft ausgeschlossen."*

Die Untersuchung von aufregenden Hypothesen an der vordersten Front der Wissenschaft erzeugt notwendigerweise eine Menge falsch-positiver Befunde, auch bei Forschung von höchster Qualität. Diese Falsch-Positiven müssen aber durch nachfolgende, kompetente Experimente wieder „ausgemerzt" werden. Reproduktion ist daher eine vornehme, hochwissenschaftliche Tätigkeit. Das Vertrackte dabei ist jedoch, dass Reproduzierbarkeit nicht einem einfachen Ja/Nein-Schema folgt.

Es ehrt den Wissenschaftler, wenn andere sich an der Reproduktion seiner Ergebnisse versuchen – denn dies bedeutet, dass sie wichtig sind. Und wenn sie nicht reproduziert werden, fängt die Wissenschaft oftmals erst richtig an, da sich dann viele Fragen stellen: Stimmte die Richtung, nur der p-Wert nicht? Was passiert, wenn man Originalexperiment und Replikation in einer Metaanalyse kombiniert? Steckt dahinter gar interessante Biologie? Oder doch eher ein bisher unerkannter Fehler? Und so weiter.

Belohnt gehören also diejenigen, deren Ergebnisse eines Reproduktionsversuchs würdig sind. Genauso wie die Wissenschaftler, die solche Experimente durchführen. Doch dies geht nur, wenn die Methoden und Ergebnisse der Studien so umfassend beschrieben werden, dass man sie auch tatsächlich nachkochen *kann*!

Die hier zitierte Literatur findet sich wie immer unter http://dirnagl.com/lj

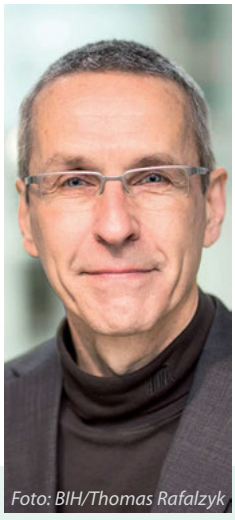

Foto: BIH/Thomas Rafalzyk

Ulrich Dirnagl

leitet die Experimentelle Neurologie an der Berliner Charité und ist Gründungsdirektor des Center for Transforming Biomedical Research am Berlin Institute of Health. Für seine Kolumne schlüpft er in die Rolle eines „Wissenschaftsnarren" – um mit Lust und Laune dem Forschungsbetrieb so manche Nase zu drehen.