Originalarbeit

Wissenschaftstheoretische Missverständnisse des BGH in Strafsachen

Überprüfung einer sogenannten„Nullhypothese“ in der Glaubhaftigkeitsbegutachtung

Jörg M. Fegert[1,2,3], Cedric Sachser[1,3], Martin Pusch[4], Andrea Kliemann[5] und Jelena Gerke[1,3]

[1] Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie, Universitätsklinikum Ulm, Deutschland
[2] Kompetenzzentrum Kinderschutz in der Medizin Baden-Württemberg, Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie, Universitätsklinikum Ulm, Deutschland
[3] Deutsches Zentrum für Psychische Gesundheit (DZPG), Mannheim, Heidelberg, Ulm, Deutschland
[4] Westpfahl Spilker Wastl, Rechtsanwälte Partnerschaft mbB, München, Deutschland
[5] Verwalterin der Professur Recht der Sozialen Dienstleistungen Fakultät I, Fach Soziale Arbeit, Universität Vechta, Deutschland

Zusammenfassung: *Hintergrund:* Der Beitrag setzt sich mit der Entscheidung des BGH in Strafsachen zur Glaubhaftigkeitsbegutachtung vom 30.7.1999 (1 StR 618/98, BGHSt 45, 164) auseinander. Hier hat der BGH in Strafsachen, auf der Basis von zwei veröffentlichten wissenschaftlichen Gutachten konkrete Anforderungen für Glaubhaftigkeitsgutachten formuliert. Diese sollten primär von der ,,Nullhypothese“ ausgehen. *Methode:* In der Auseinandersetzung mit Originalzitaten sollen sich widersprechende wissenschaftstheoretische Postulate in den Expertengutachten und die Rezeption dieser Prinzipien im BGH-Urteil analysiert werden. *Ergebnisse:* Angesichts der zentralen Bedeutung dieser BGH-Entscheidung werden die Originalgutachten auf ihre wissenschaftstheoretischen Inhalte hin analysiert. Als wissenschaftliche Methodik hat der BGH das Ausgehen von der Annahme, die Aussage sei unwahr – sogenannte „Nullhypothese“ – formuliert. Er bezog sich dabei auf Poppers Deduktionismus, ohne aber auf Regeln der Hypothesenüberprüfung einzugehen. Basierend auf dem zweiten Gutachten, welches mit der induktiven statistischen Aggregation in Bezug auf die Wahrheitsfindung argumentiert, übernimmt der BGH die Annahme, dass „durch das Zusammenwirken der Indikatoren deren Fehleranteile insgesamt gesenkt“ würden. Diesem Umstand liege das mathematische und psychometrisch eingehend untersuchte Prinzip der Aggregation zugrunde. Kausal verknüpfend behauptet der BGH: „Dementsprechend lagen die mit Realkennzeichen in Forschungsvorhaben erzielten Ergebnisse regelmäßig deutlich über dem Zufallsniveau. Allerdings bestanden dabei teilweise nicht unerhebliche Fehlerspannen.“ *Fazit:* Die mehr als 25 Jahre alte Entscheidung hat die wissenschaftliche Weiterentwicklung der Aussageanalyse und die Überprüfung ihrer Verwendung, z. B. in unterschiedlichen Altersgruppen oder bei Personen mit erheblichen Beeinträchtigungen, eher gehemmt als gefördert. Von Betroffenen wird die Grundannahme, zunächst gelte die Aussage als unwahr, als epistemische Ungerechtigkeit wahrgenommen. Auch eines der dem BGH vorliegenden Gutachten betont die Bedeutung der Ausgangsfragestellung mit Blick auf Bestätigungstendenzen der Ausgangshypothese. Der Beitrag stellt dar, wie ein Jargon der Wissenschaftstheorie einem Vorgehen mit erheblichen Limitationen eine fast dogmatische Anwendungspraxis sichern konnte, die weit über den Anwendungsbereich des Strafrechts, auf den die ursprünglich berechtigte Orientierung am Zweifelsgrundsatz abgestellt war, hinausragt.

Schlüsselwörter: Hypothesenüberprüfung, wissenschaftliche Methoden, induktiv statistische Aggregation, Deduktion, Positivismus

Epistemological Misunderstandings of the German Federal Court of Justice in Criminal Cases Regarding the Null Hypothesis: Verification in the Credibility Assessment

Abstract: *Background:* The article deals with the decision of the German Federal Court of Justice (Bundesgerichtshof, BGH) in criminal matters regarding credibility assessment dated 30 July 1999 (1 StR 618/98, BGHSt 45, 164). Regarding criminal matters, the BGH formulated specific requirements for credibility assessments based on two published scientific expert reports. *Method:* We analyzed conflicting postulates of scientific theory in the expert reports and the reception of these principles in the BGH judgment by examining the original quotes. *Results:* Given the central importance of this BGH decision, we analyzed the original expert reports for their epistemological content. The BGH formulated the scientific approach of starting from the assumption that the statement is untrue – the so-called “null hypothesis”. In doing so, it referred to Popper’s deductivism, albeit without addressing the rules of hypothesis testing. Based on the second expert report, which argues for inductive

Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie (2024), *52* (6), 342–352
https://doi.org/10.1024/1422-4917/a000995

statistical aggregation concerning truth findings, the BGH adopts the assumption that "based on the interaction of the indicators, their error rates would be reduced overall." This approach goes back to the mathematical and psychometrically thoroughly investigated principle of aggregation. Applying causally linking, the BGH asserts: "Accordingly, the results obtained with real characteristics in research projects regularly exceeded the random level. However, there were sometimes considerable margins of error." *Conclusion:* This decision, declared more than 25 years ago, has rather hindered than promoted the scientific advancement of statement analysis and the examination of its use, e.g., in different age groups or individuals with significant impairments. The basic assumption that the statement should be considered untrue initially is perceived by those affected as an epistemic injustice. One of the expert reports submitted to the BGH also emphasizes the importance of the initial questioning concerning confirmation tendencies of the initial hypothesis. The question is how jargon from scientific theory could secure an almost dogmatic application practice with considerable limitations that extends far beyond the scope of criminal law – for which the original orientation was based on the principle of doubt.

Keywords: hypothesis testing, scientific methods, inductive statistical aggregation, deduction, positivism

Karlsruhe locuta – causa finita

In seiner Entscheidung vom 30.7.1999 (1 StR 618/98, BGHSt 45, 164) hat der BGH in Strafsachen auf der Basis zweier veröffentlichter wissenschaftlicher Gutachten konkrete Anforderungen für aussagepsychologische Glaubhaftigkeitsgutachten formuliert. Das umstrittene Vorgutachten, dessen Anfechtung die erkennende Strafkammer zunächst abgelehnt hatte, weil die Sachkunde der sorgfältigen und forensisch erfahrenen Sachverständigen außer Zweifel stehe, genüge methodischen Mindeststandards nicht. Grundsätzlich stellt der BGH fest, dass die angewandten Methoden dem jeweils anerkannten wissenschaftlichen Kenntnisstand gerecht werden müssen. Wörtlich spricht der BGH dabei von „Test- und Untersuchungsverfahren": „Die eingesetzten Test- und Untersuchungsverfahren müssen zudem durch die gebildeten Hypothesen indiziert, d.h. geeignet sein, zu deren Überprüfung beizutragen." Dabei ist festzuhalten, dass im Gegensatz zu psychodiagnostischen Tests, das empfohlene Vorgehen bei der Überprüfung der Unwahr-Hypothese kein standardisiertes, normiertes Verfahren ist.[1]

Der Beitrag soll mit Blick auf die wissenschaftstheoretischen Postulate in den Expertengutachten die Rezeption dieser Prinzipien im BGH-Urteil analysieren (ausführlich in Fegert et al., 2024). Untersucht werden soll auch die mit bestimmten wissenschaftstheoretischen Begrifflichkeiten wie „Hypothesenüberprüfung" oder Konzepten, etwa dem des Fallibilismus[2], verbundene rhetorische Aufladung im Sinne einer Inszenierung von Wissenschaftlichkeit („scientifically based standards"; Niehaus & Krause, 2023a) oder „Wissenschaftsorientierung" (Niehaus & Krause, 2023b), was immer das sein mag. Dies ist nicht nur von rechtshistorischem Interesse. Das Label „Wissenschaftsorientierung" ist in der Debatte um Probleme in der Glaubhaftigkeitsbegutachtung zu einem Kampfbegriff einer Gruppe von klinisch in der Regel nicht ausgebildeten Rechtspsycholog_innen geworden, welche Kritiker_innen des im deutschsprachigen Raum üblichen Vorgehens grundsätzlich vorwerfen, den Boden wissenschaftlicher Methoden zu verlassen (z.B. Niehaus & Kraus, 2023a, 2023b). Auffallend ist, dass seit dem BGH-Urteil am Ende des letzten Jahrhunderts deutschsprachige empirische Forschung zur Glaubhaftigkeitsbegutachtung die dort angesprochenen Wissenslücken nicht füllen konnte. Seit dem erfolgreichen Förderschwerpunkt „Recht und Verhalten" der Volkswagenstiftung, wurden keine interdisziplinären Forschungsschwerpunkte zu den Tatsachengrundlagen und juristischen Konsequenzen mehr ausgeschrieben. So beendete das BGH-Urteil, in dem selbst noch erheblicher Forschungsbedarf in Bezug auf die Anwendbarkeit der Methode bei bestimmten Populationen angemahnt wurde, nicht nur eine wissenschaftliche Debatte um ein geeignetes Vorgehen, sondern führte zu einer dogmatischen Festlegung. durch das „Adeln" einer bestimmten Vorgehensweise mit einem höchstrichterlichen Gütesiegel unter Berufung auf keinen Geringeren als Popper.

Das Vorgehen des BGH widerspricht, wie zu zeigen sein wird, gerade Poppers Wissenschaftsverständnis, welches er in seinem Buch „Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft" 1935 dargestellt hat. Er grenzt sich dabei vom logischen Positivismus des sogenannten Wiener Kreises ab, der in der Tradition der bisherigen Erkenntnistheorie den Wahrheitsanspruch

[1] Roma locuta – causa finita bezeichnet sprichwörtlich die kirchenrechtliche Möglichkeit des Papstes, Debatten durch ein Machtwort zu beenden. Rom hat gesprochen und die Sache ist erledigt. Übertragen wird dies in dieser Überschrift nun angewandt, um deutlich zu machen, dass das BGH-Urteil vom 30.7.1999 nicht nur Ordnung geschaffen hat und gewisse Minimalstandards beschrieben hat (vgl. König & Fegert 2009), sondern dass damit eine bestimmte Vorgehensweise zementiert wurde und eine weitere wissenschaftliche Auseinandersetzung weitgehend ausgeblieben ist.

[2] Erkenntnistheoretische Position, nach der es in der Erkenntnis keine Gewissheit geben kann, ob eine Aussage wahr oder falsch ist (vgl. Dorsch, Lexikon der Psychologie. Abgerufen am 16.5.2024 unter https://dorsch.hogrefe.com/stichwort/wissenschaftstheorie).

wissenschaftlicher Gesetze induktiv durch empirische Einzelbeobachtungen untermauern wollte. Dagegen führt Popper an, dass ein Wahrheitsbeweis durch einzelne Beobachtungen induktiv logisch nicht möglich sei. Geboten sei die deduktive Falsifizierung oder Bestätigung durch Versuch und Irrtum. Im Kern bedeutet Poppers Wissenschaftsverständnis, dass eine Theorie oder Regel (Axiom) in einem empirisch wissenschaftlichen System durch die Realität, also in der Anwendung, überprüfbar sein müsse. Dies bezeichnet er als Falsifizierbarkeit. Sein Vorgehen bedeutet, dass postulierte Gesetzmäßigkeiten oder Hypothesen kritisch geprüft werden müssen. Theorien oder Hypothesen, die eine kritische Prüfung bestehen können, gelten bis zu ihrer Widerlegung als bewährt. Theoretische Annahmen sollen also immer wieder durch neue, bessere ersetzt werden, damit sie schließlich der Wahrheit im Sinne einer Übereinstimmung mit der Wirklichkeit immer näherkommen. Letztendlich gibt es aber nie eine absolute Gewissheit darüber, ob eine Aussage wahr oder falsch ist.

> So beschreibt der BGH das Vorgehen:
> „Der Sachverständige nimmt daher bei der Begutachtung zunächst an, die Aussage sei unwahr (sogenannte Null-Hypothese). Zur Prüfung dieser Annahme hat er weitere Hypothesen zu bilden. Ergibt seine Prüfstrategie, dass die Unwahr-Hypothesen mit den erhobenen Fakten nicht mehr in Übereinstimmung stehen kann, so wird sie verworfen und es gilt dann die Alternativ-Hypothese, dass es sich um eine wahre Aussage handelt. Die Bildung relevanter Hypothesen ist daher von ausschlaggebender Bedeutung für Inhalt und (methodischen) Ablauf einer Glaubhaftigkeitsbegutachtung. Sie stellt nach wissenschaftlichen Prinzipien einen wesentlichen, unerlässlichen Teil des Begutachtungsprozesses dar (...)".

Karl Popper geht davon aus, dass eine Annahme niemals bewiesen, sondern nur widerlegt werden könne. Übersetzt man diese Formulierungen des BGH in eine generelle Annahme, so würde dies bedeuten: Solange nicht ausreichende Fakten erhoben werden können, gilt jede Aussage als falsch. Dieser Satz lässt sich tatsächlich einfach falsifizieren und taugt deshalb nicht als grundlegende Annahme, da es selbstverständlich Aussagen gibt, welche richtig sind, bei denen aber im Rahmen der Überprüfung einzelner Hypothesen wie z. B. der Suggestionshypothese Restzweifel geblieben sind. Niemand würde eine solche Aussage als deterministisches Gesetz bezeichnen wollen. Regeln daraus abzuleiten, ist gewagt. Denn nach Poppers Bild vom schwarzen Schwan reicht eine einzige der Regel widersprechende Beobachtung, um das ganze Konstrukt zu Fall zu bringen. Wenn es also Aussagen gibt, bei denen Restzweifel bleiben, weshalb die Unwahrhypothese nicht verworfen werden kann, die aber dennoch tatsächlich wahr sind, dann muss die zugrunde liegende Regel als falsifiziert verworfen werden. Tatsächlich gibt es aber in der Wissenschaft keine absolute Gewissheit über ja oder nein. Vielmehr können gerade in Bezug auf psychologische, psychopathologische Variablen und Variablen des Sozialverhaltens eher fundierte und z. B. bei normierten Testverfahren quantifizierbare Wahrscheinlichkeitsbewertungen abgegeben werden.

Semper reformanda?

Wenn schon einige Theolog_innen von der Kirche als einer stets zu Reformierenden sprechen, um wie viel mehr gilt dies für die Life Sciences und generell für empirische Wissenschaft, in der typischerweise ständig neuere Erkenntnisse ältere weiterentwickeln oder präzisieren, manchmal aber auch revidieren. Gerade im Zusammenhang mit der evidenzbasierten Medizin macht die Metapher der sinkenden Halbwertszeit des Wissens die Runde (Sauerland & Waffenschmidt, 2018). Der medizinische Fakultätentag (o. D.) zitiert in einer offiziellen Stellungnahme zur Vermittlung von Wissenschaftskompetenz McDeavitt, 2014 (zit. aus Medizinischer Fakultätentag, o. D.), der damals davon ausging, dass das medizinische Wissen sich in unserer Zeitperiode nach 2020 alle 73 Tage verdoppelt. Man kann sich über solche pseudopräzisen Angaben wundern, unbestritten ist aber, dass eine rasche Steigerung der Original-Publikationstätigkeit für fast alle wissenschaftlichen Felder typisch ist. Für die Life Sciences kann man hier durchaus von einem exponentiellen Wachstum sprechen. Während die letzten 25 Jahre nicht zuletzt durch potenziell traumatisierende Ereignisse wie 9/11, Hurrikan Katrina oder die schreckliche Tat-Serie in Utøya zentrale Fortschritte in der Erforschung der Traumareaktionen und in Bezug auf evidenzbasierte Traumatherapien brachten, wobei immer wieder neue Methoden und Zugänge publiziert, diskutiert und überprüft wurden, wird im Feld der Begutachtung von Betroffenen in Strafverfahren, aber auch in anderen Verfahren wie z. B. im Sozialen Entschädigungsrecht bei der Tatfeststellung auf die vom BGH sanktionierte weitgehend unveränderte aussagepsychologische Methodik der sogenannten Hypothesenüberprüfung zurückgegriffen. Hier werden unterschiedliche Annahmen formuliert, welche durch eine Aussageanalyse und eine gründliche Untersuchung der Entstehungsbedingungen einer Aussage „überprüft" werden sollen. Zum Zeitpunkt des BGH-Urteils spielte die Aussageanalyse mit sogenannten „Realkennzeichen" (Steller & Köhnken, 1989) eine zentrale Rolle. Zugrunde lag die Annahme, dass Lügen oder Falschaussagen sich aufgrund bestimmter Merk-

male von erlebnisbasierten Aussagen unterscheiden. Da die Wahrheit einer Aussage nicht wissenschaftlich überprüft werden könne, wenn keine anderen Beweise oder Außenkriterien vorliegen, wurde für das Vorgehen der Versuch der Widerlegung der Unwahrhypothese oder „Nullhypothese" (so der vom BGH aufgegriffene Begriff) gewählt. Zunächst gilt dabei die Aussage als unwahr, bis durch die Überprüfung einzelner Merkmale und Charakteristika kein Restzweifel mehr bleibt.

Von Betroffenen ist diese theoretische Unterstellung der Unwahrheit ihrer Aussagen wiederholt prinzipiell kritisiert worden (Fegert, Gerke & Rassenhofer, 2018). Die Ausformulierung der Unwahrhypothese bei der Herangehensweise einer kriterienorientierten Aussageanalyse wird heute in der Debatte auch von Rechtspsycholog_innen in der Regel nicht mehr häufig benützt, sondern man bezieht sich gemeinhin eher auf die deutlich adäquatere, z.B. von Renate Volbert schon vor der BGH-Entscheidung so ausformulierte Fragestellung: "Könnte dieses Kind unter den gegebenen individuellen Voraussetzungen unter den gegebenen Befragungsumständen und unter Berücksichtigung der im konkreten Fall möglichen Einflüsse von dritten diese spezifische Aussage machen, ohne daß sie auf einem realen Erlebnis Hintergrund basiert?" (Volbert, 1995). Damit verschiebt sich auch ein gewisser Fokus hin zur Klärung möglicher suggestiver Einflüsse im Kontext der Aussage-Entstehung. Könnte durch Befragungsmängel, durch Anschauungsmaterial z.B. im Sexualkundeunterricht oder gar durch eine Traumabearbeitung im Rahmen einer Psychotherapie die Aussage suggeriert worden sein. Da empirisch nachgewiesen ist, dass suggerierte Aussagen aufgrund von Aussage-Merkmalen in der Aussage-Analyse nicht unterschieden werden können (Laney & Takarangi, 2013; Shaw, 2020), kann auch die scheinbare wissenschaftliche Methode mit der Betrachtung von Merkmalsunterschieden keine Aussage bringen. Deshalb hat die Überprüfung der sogenannten Suggestionshypothese in den letzten Jahren in der Praxis eine immer stärkere Bedeutung bekommen. Allerdings lässt sich hierauf keinesfalls die Argumentation anwenden, dass durch Aggregation von Beobachtungen aus „schwachen Indikatoren solche mit Indizcharakter" würden.

Hypothesenüberprüfung ohne Hypothesentest

„Kann denn ein Götze einen guten Rat erteilen? Er ist mit Gold und Silber überzogen, aber er hat kein Leben in sich!" (Habakuk 2,19b)

Die Wortwahl „Hypothesenüberprüfung" will den Glanz des Goldes wissenschaftlicher empirischer Überprüfung suggerieren. Doch was heißt Hypothesenüberprüfung in diesem Zusammenhang? Gemeint ist die mehr oder weniger systematische Suche nach Inkongruenzen in Aussagen oder z.B. möglichen suggestiven Einflüssen – sei es im Rahmen der Erstaussage oder durch Befragungspersonen oder aber durch psychotherapeutische Interventionen etc. Auch mögliche mediale Beeinflussungen oder Wissen aus dem Aufklärungsunterricht und andere mögliche Ereignisse des alltäglichen Lebens, können hier sozusagen „das Haar in der Suppe" bilden, welches die Begutachtenden zu dem Schluss führt, dass suggestive Einflüsse nicht ausgeschlossen werden können. Es reicht letzten Endes ein einziger Zweifel, um am Abschluss eines Gutachtens zu der Einschätzung zu kommen, dass nicht ausgeschlossen werden kann, dass diese Aussage auch ohne reales Erleben hätte zustande kommen können. Dies bedeutet natürlich nicht, dass unterstellt wird, dass die Aussage unwahr sei, sondern nur, dass ein gewisser Restzweifel bleibt, de facto kann aber die Unwahr- oder „Nullhypothese" nicht widerlegt werden und deshalb erfolgt in der Regel, wenn nicht andere Beweise vorliegen, „im Zweifel für die/den Angeklagte_n" eine entsprechende Entscheidung des Gerichts.

Bei der Tagung zu Therapie- und Strafverfahren im BMJ am 6. und 7.10.2022 bezog sich auch eine in einem Strafsenat tätige Richterin am BGH in ihrem einleitenden Referat direkt auf Popper und leitete von ihm die wissenschaftliche Fundierung der Anwendung der sogenannten „Nullhypothese" in der Glaubhaftigkeitsbegutachtung ab. Auftaktreferat und das Co-Referat von Adorno bei der Arbeitstagung der Deutschen Gesellschaft für Soziologie im Oktober 1961 in Tübingen zur Logik der Sozialwissenschaften eröffnete eine zentrale wissenschaftstheoretische Debatte über die methodische Herangehensweise in den Sozialwissenschaften, bekannt als sog. „Positivismusstreit" (Adorno, 1969). Popper vertrat hier den Ansatz des kritischen Rationalismus. Er grenzte sich von dem Prinzip der Wertfreiheit der Wissenschaft, wie es noch der Soziologe Max Weber postulierte, ab und stellte übrigens in Übereinstimmung mit Adorno fest, dass Werturteile bei Theorienbildung in der Wissenschaft immer eine Rolle spielen. Popper postulierte, dass das wissenschaftliche Prinzip sich in den Natur- und Sozialwissenschaften gleiche. Es gehe darum, „Lösungsversuche" auszuprobieren. Er sprach sich, wie bereits beschrieben, *gegen eine induktive Theoriebildung* aus, also gegen ein Sammeln von Einzelbeobachtungen, und plädierte für die *deduktive* ***Überprüfung*** *oder Testung* von wissenschaftlichen Gesetzmäßigkeiten oder Problemlösungsversuchen. Die deduktive Wissenschaftsmethode war schon von Aristoteles dem induktiven Vorgehen gegenübergestellt worden, d.h.

der Gewinnung einer Gesetzmäßigkeit oder Allgemeinaussage aus der Betrachtung mehrerer Einzelfälle. Die Deduktion entsprach der Überprüfung einer allgemeinen Regel durch empirische spezielle Tatsachen. Eine Spezialform der Deduktion durch empirische Beobachtungen ist der sogenannte „Hypothetico-Deduktivismus". Nach dieser wissenschaftlichen Methode werden Hypothesen formuliert, um diese durch Beobachtungen, die den deduktiven Konsequenzen der Hypothesen zuwiderlaufen, zu falsifizieren. Die wissenschaftliche Hypothesen-Prüfung beginnt mit der richtigen Formulierung einer Hypothese, die eine Widerlegbarkeit ebenso wie Operationalisierbarkeit ermöglichen muss. Eine wissenschaftliche Untersuchung oder Hypothesen-Testung ist eine Beobachtung, welche in einem Experiment systematisch durchgeführt wird, um letztendlich nach bestimmten Regeln eine Entscheidung über die Beibehaltung bzw. die Ablehnung einer Hypothese treffen zu können. Für solche Verfahren gelten die üblichen Test-Gütekriterien der Objektivität, Reliabilität und Validität. Normalerweise wird bei der wissenschaftlichen Hypothesen-Überprüfung eine „Nullhypothese" und eine Alternativ-Hypothese vor der Untersuchung festgelegt und ein Auswertungsplan definiert, der aufgrund einer Entscheidungsregel das Verwerfen bzw. Nichtverwerfen der „Nullhypothese" vorsieht. In der quantitativen experimentellen Hypothesen-Überprüfung werden Signifikanz-Tests zur Überprüfung der Richtigkeit der „Nullhypothese" eingesetzt. Solche gruppenstatistischen Verfahren werden z. B. verwendet, um wissenschaftliche Tests oder Untersuchungsmethoden zu überprüfen.

Während das eine der BGH-Entscheidung zugrunde liegende Gutachten von Steller und Volbert (1999) die damalige Methodik forensisch aussagepsychologischer Begutachtung im Sinne der Autor_innen beschrieb und dabei auch auf Limitationen (S. 65 ff.) und die insgesamt wenig erfolgreiche wissenschaftliche Überprüfung in empirischen statistischen Untersuchungen (S. 77 f.) einging, bezog sich das Gutachten von Fiedler und Schmid (1999) auf die Methodik. In ihrem Gutachten unterscheiden sie wissenschaftstheoretisch deduktiv-nomologische Beweise einerseits von induktiv-statistischen Schlüssen andererseits. Die statistische Überprüfung der Treffsicherheit einzelner Glaubhaftigkeitskriterien entsprach einem solchen deduktiven Vorgehen, in dem überprüft wurde, ob aufgrund einer bestimmten Gesetzmäßigkeit, also hier dem ausformulierten Kriterium z. B. Detailreichtum, wahre von konfabulierten Aussagen unterschieden werden können. Fiedler und Schmid (1999) bezeichnen die Annahme, dass universell anwendbare Gesetze in diesem Kontext der Begutachtung existieren könnten, als unrealistisch. Gleichzeitig sehen sie aber in einzelnen schwachen Merkmalen „Indikatoren, die für sich genommen alle nur von bescheidener Aussagekraft sind, obwohl sie im Erwartungswert (Durchschnitt) besser als der Zufall sein müssen, dann könne durch die statistische Aggregation die Treffsicherheit deutlich gebessert werden". Durch das Prinzip der Aggregation werde dann aus vielen schwachen Indikatoren eine robuste Schlussfolgerung mit einem beträchtlichen diagnostischen Wert. Fehleranteile der einzelnen imperfekten Gesetzmäßigkeiten seien per definitionem statistisch unabhängig und würden sich daher gegenseitig eher neutralisieren, während die verlässlichen Anteile eine Gemeinsamkeit aufwiesen, nämlich die zugrunde liegende Größe, also die tatsächliche Wahrheit der Aussage. Ganz zentral ist eine Feststellung der Autoren dieses Gutachtens zu Fehlschlüssen durch selektive Nutzung von Indikatoren:

> Eine entscheidende Voraussetzung für diagnostische Nutzung solcher Indikator-Systeme – und mitverantwortlich für die empirische mehrfach beobachtete Genauigkeit solcher Systeme (Ambady & Rosenthal, 1992) – ist jedoch wie bereits oben klargestellt die repräsentative, nicht-selektive Auswahl der Indikatoren. Typisch für die Bedingungen, unter denen die Diskrimination von wahren und falschen Aussagen aufgrund minimaler Information erfolgreich war, ist die Nicht-Selektivität der beurteilten Beobachtungen (vgl. Bunswik's, 1955, Forderung nach ‚representative sampling'). Durch Einschränkung der Information auf wenige selektive Indikatoren, die einem bestimmten favorisierten Modell entsprechen, und Ignorieren anderer Indikatoren, die andere denkbare Modelle bestätigen könnten, werden unter Umständen erhebliche Fehler erzeugt. So zeigen unmittelbar mit Glaubwürdigkeit befasste Experimente (z. B. Zuckerman, Koestner, Colella, & Alton, 1984), dass Aussagen eher für falsch gehalten werden, *wenn Urteiler die Hypothese einer möglichen Lüge testen* (Hervorhebung durch die Verf.), während dieselben Aussagen eher für wahr gehalten werden, wenn die Hypothese einer wahren Äußerung focussiert wird. In der psychologischen Forschung im allgemeinen (Jussim, 1991; Koehler, 1991) und der Forschung zum Hypothesentesten in Gesprächen und Interviews im Besonderen wurde vielfach demonstriert, dass die Ergebnisse systematisch in Richtung auf die Ausgangshypothese verzerrt sind (Snyder, 1984; Pyszczynski & Greenberg, 1988; Tversky & Kahneman, 1974; Zuckerman et al., 1995). Einer von mehreren Gründen für diesen sogenannten „confirmation bias" (Snyder & Swann, 1978) bzw. „auto-verification effect" (Fiedler, Walther & Nickel, 1999) ist die einseitige, nicht-repräsentative Suche nach Indikatoren für die leitende Hypothese und die gleichzeitige Vernachlässigung von Indikatoren für alternative Hypothesen. (Fiedler & Schmid, 1999).

Die vom BGH als Ausgangshypothese gewählte Unwahrannahme verfälscht also das Ergebnis der Begutachtung nach diesen Überlegungen des BGH-Gutachtens von Fiedler und Schmid dadurch, dass hauptsächlich Belege für ein Zustandekommen der Aussage ohne eine Realitätsbasierung gesucht werden bzw. wie es der BGH sehr vereinfacht formulierte, dass die Aussage zunächst als falsch gilt. Das von Fiedler und Schmid beschriebene Prinzip der Datenaggregation zur Reduktion des Messfehlers von Einzeldaten und zur Erhaltung verallgemeinerbarer Resultate wird, dieser Eindruck drängt sich auf, vom BGH intentional oder unabsichtlich fehlinterpretiert oder komplett missverstanden, denn die Gutachtenden beschreiben ein induktives Vorgehen einer Gesamtschau, nicht eine deduktive Überprüfung. Wörtlich heißt es im Urteil des BGH:

> Zur Durchführung der Analyse der Aussagequalität sind auf der Basis der dargestellten Annahmen Merkmale zusammengestellt worden, denen indizielle Bedeutung für die Entscheidung zukommen kann, ob die Angaben der untersuchten Person auf tatsächlichem Erleben beruhen. Es handelt sich um aussageimmanente Qualitätsmerkmale (z. B. logische Konsistenz, quantitativer Detailreichtum, raumzeitliche Verknüpfung, Schilderungen ausgefallener Einzelheiten und psychischer Vorgänge, Entlastung des Beschuldigten, deliktspezifische Aussage-Elemente), deren Auftreten in einer Aussage als Hinweis auf die Glaubhaftigkeit der Angaben gilt […]. Diese sog. Realkennzeichen können als grundsätzlich empirisch überprüft angesehen werden. Zwar handelt es sich um Indikatoren mit jeweils für sich genommen nur geringer Validität, d. h. mit durchschnittlich nur wenig über dem Zufallsniveau liegender Bedeutung. Eine gutachterliche Schlussfolgerung kann aber eine beträchtlich höhere Aussagekraft und damit Indizwert für die Glaubhaftigkeit zu beurteilender Angaben erlangen, wenn sie aus der Gesamtheit aller Indikatoren abgeleitet wird. Denn durch das Zusammenwirken der Indikatoren werden deren Fehleranteile insgesamt gesenkt. Diesem Umstand liegt das mathematisch und psychometrisch eingehend untersuchte Prinzip der Aggregation zugrunde (Gutachten Prof. Dr. Fiedler). Dementsprechend lagen die mit Realkennzeichen in Forschungsvorhaben erzielten Ergebnisse regelmäßig deutlich über dem Zufallsniveau. Allerdings bestanden dabei teilweise nicht unerhebliche Fehlerspannen. Inwieweit ihre Bedeutung bei Verwendung gegenüber Personen aus unterschiedlichen Altersgruppen differieren kann, ist völlig offen.

Beim gebräuchlichen Vorgehen im Rahmen der Glaubhaftigkeitsbegutachtung kommt es aber nicht zur Anwendung einer „induktiv-statistischen Methode“, wo viele systematische und nicht durch eine einseitige Fragestellung selegierte Einzelbeobachtungen dazu führen, eine relativ sichere Gesamtbewertung abgeben zu können, sondern eine Einzelbeobachtung, ein Restzweifel, z. B. ein möglicher suggestiver Einfluss, kann dazu führen, dass wegen dieses einen Merkmals nicht ausgeschlossen werden kann, dass die entsprechende Aussage ohne eine reale Erlebnisgrundlage zustande gekommen ist. Dies widerspricht dem Prinzip des „vicarious functioning“, welches im Gutachten Fiedler und Schmid (1999, S. 22) an einem Beispiel aus der Optik illustriert wurde:

> Dieser als ‚vicarious functioning‘ bezeichnete Vorteil findet sich übrigens nicht nur in diagnostischen Modellen, sondern auch in vielen natürlichen Systemen, die unter Unsicherheit Lösungen finden und Entscheidungen treffen müssen, deren Effizienz angesichts der Schwäche der verwendeten Indikatoren überraschend hoch ist (Brunswik, 1955; Gigerenzer & Goldstein, 1996). Ein Beispiel ist etwa menschliches Tiefensehen (Entfernungssehen), wo für sich genommen schwache Indikatoren (Glanz der Oberfläche, Disparität der beiden Netzhautbilder etc.) zusammen erstaunliche Genauigkeit erzielen und den Ausfall einzelner Indikatoren leicht verkraften können. Diese Bezüge seien hier nur deshalb erwähnt, um deutlich zu machen, dass ein psychologischer und mathematischer Bezugsrahmen zur Erklärung der erstaunlichen Genauigkeit von Systemen schwacher Prädiktoren schon seit langem existiert und formal sehr weit entwickelt ist.

Um in diesem Beispiel zu bleiben, fokussiert die Glaubhaftigkeitsbegutachtung auf den Ausfall einzelner Indikatoren und entspricht damit eben nicht der erstaunlichen Genauigkeit der Gesamtwahrnehmung beim Sehen, da eine Aussage-Analyse mit Gewichtung auf die Gesamtwahrnehmung unterbleibt. Einzelne Restzweifel entsprechend punktuellen Wahrnehmungsbeeinträchtigungen beim Beispiel des Sehens, werden so nicht ausgeglichen, sondern drohen in einer binären Entscheidung die komplette Restwahrnehmung auszublenden.

„Judex non calculat“ – Der Richter rechnet nicht

Ein altes juristisches Sprichwort soll darlegen, dass die Rechtswissenschaft eine Textwissenschaft ist, welche

nicht primär durch die Kenntnis und Anwendung mathematischer oder naturwissenschaftlicher Regeln geprägt ist. Wie alle generellen Sätze lässt auch dieser sich schnell falsifizieren, denke man nur in familienrechtlichen Verfahren an den Versorgungsausgleich o.Ä. Gleichwohl weist dieses Sprichwort in Bezug auf den Gegenstand dieses Beitrags, nämlich die argumentative Bedeutung von Wissenschaftstheorie, auf ein Phänomen hin, nämlich dass den Jurist_innen am BGH die statistischen Regeln und Begriffsbedeutungen, welche sich hinter Begriffen wie „deduktives Vorgehen der Hypothesen-Überprüfung", „Verwerfen oder Beibehalten einer „Nullhypothese", „induktive statistische Aggregation" etc. nicht bekannt sein mussten. Vielmehr bediente man sich offensichtlich in den beiden Gutachten wie in einem Steinbruch und konstruierte daraus quasi eine Beweisregel, welche in den Fällen, in denen Gutachtende beteiligt werden, aus dem Gerichtsverfahren ausgelagert im Vorfeld angewendet wird und regelmäßig zu einer binären (ja/nein) weitgehend vorentscheidenden Bedeutung gelangt. Klare statistische Methoden und Schwellen, wann eine „Nullhypothese" verworfen werden muss oder wann eine Alternativ-Hypothese gilt, werden nicht betrachtet und können auch gar nicht angegeben werden, da die Methode der Aussage-Analyse kein standardisiertes Testverfahren ist, sondern eine exploratorische Textanalyse mit Blick auf unterschiedliche Ausgangsannahmen im Sinne der Konstanzanalyse, der Analyse möglicher suggestiver Bedingungen, die die Aussagegenese begleitet haben könnten etc. Durch die Verwendung der Terminologie des statistischen Testens in einem völlig anderen Kontext kreiert der BGH einen Sound, der eben nicht dem Verständnis des methodischen Vorgehens dient, sondern eher eine vermeintliche Vertrautheit mit den wissenschaftstheoretischen Implikationen signalisieren soll, ohne dass dies aber tatsächlich der Fall wäre. Es liegt im Wesen (nicht nur) der richterlichen Beweiswürdigung, dass es keine absolute Gewissheit betreffend die zu beurteilenden Sachverhalte geben kann. Selbst ein Test oder ein Geständnis sind nicht in der Lage, diese zu liefern. Eine Vorgehensweise, die allerdings die Würdigung komplexer Irrtumswahrscheinlichkeiten auf eine binäre Vorentscheidung reduziert, widerspricht diametral dem Wesen und der Vorgehensweise richterlicher Überzeugungsbildung.

Tatsächlich sind die statistischen Regeln des Hypothesen-Überprüfens voraussetzungsvoll und setzen sehr viel Wissen über geeignete Methoden und Bedingungen in einem experimentellen oder quasi experimentellen Erhebungsdesign voraus. Der BGH nutzt die Begriffe als Metaphern der Wissenschaftlichkeit und suggeriert somit quasi im Plauderton eine wissenschaftliche Durchdringung. Wahllos wird ein Test-Gütekriterium, nämlich die Validität in Bezug auf die kriterienorientierte Aussageanalyse, angesprochen. Zu anderen Gütekriterien Reliabilität, insbesondere Interrater-Reliabilität, Test-Retest-Reliabilität, Objektivität der Durchführung etc. wird nichts gesagt. Generell wird festgestellt, dass die einzelnen Kriterien der Aussageanalyse, z.B. in der Konstanz-Analyse, nur geringe Validität hätten. Gemeint wird vom BGH damit, dass die Ergebnisse der experimentellen Überprüfung der Diskriminationsfähigkeit dieser Merkmale in Bezug auf wahre und erfundene Aussagen meist in Studierenden-Populationen nur wenig über der Zufallswahrscheinlichkeit lagen. Assoziativ scheint im oben angegebenen Zitat aus dem BGH-Urteil auf, dass es den Richter_innen vermutlich bewusst war, dass es eine Bedeutung hat, in welcher Stichprobe ein Phänomen untersucht wurde, also dass entsprechende Aussagen nur für entsprechende Grundgesamtheiten gelten (Bayes' Theorem bedingter Wahrscheinlichkeiten). Im Urteil wird z.B. angesprochen, dass man nicht genügend über Altersunterschiede wisse; das Gleiche gilt für andere Subpopulationen wie z.B. Menschen, die in ihren Aussage-Möglichkeiten beeinträchtigt sind, wie Menschen mit geistiger Behinderung etc.

Wie kommt man also vom Stroh der Einzelmerkmale zum Gold einer quasi Beweisregel? Indem man aus dem Methoden-Gutachten an anderer Stelle die statistisch-induktive Methode der Aggregation anspricht. Wen schert es da, dass man vorher wissenschaftstheoretisch induktive Schlüsse abgelehnt hat, sondern auf deduktiver Falsifikation beharrt hat? Muss es irritieren, dass keine Statistik einzelner Merkmale nach wissenschaftlichen Regeln analysiert wird? Aus dem Stroh der Einzelmerkmale wird durch diese Amalgamierung wissenschaftstheoretisch unvereinbarer Argumente eine scheinbar wissenschaftlich belegte Methode zur Feststellung über die Wahrheit (sic) einer Aussage. Dass vom Gutachtenden gewählte Beispiel des Sehens macht deutlich, wie sehr hier das Grundprinzip dieser statistisch-induktiven Argumentation missverstanden wurde. Die Gutachter Fiedler und Schmid gehen davon aus, dass aus der Vielzahl erhobener Einzelkriterien, von denen man weiß, dass sie kaum zwischen Wahrheit und Unwahrheit einer Aussage unterscheiden können, eine robustere Aussage dadurch entsteht, indem alle Merkmale, die tatsächlich das Richtige repräsentieren, statistisch miteinander verbunden sind und damit zu einem richtigen Gesamteindruck führen. Es geht um eine statistische Aggregation, also eine Zusammenfassung vieler Variablen, die divergierende Einzelbeobachtungen unter den Tisch fallen lässt, um induktiv zu einem insgesamt schlüssigen und plausiblen Gesamtbild (welches der Wahrheit nahekommt) zu kommen. Was macht aber die Aussagepsychologie in der Glaubhaftigkeitsbegutachtung? Es erfolgt keine statisti-

sche Aggregation von Einzelmerkmalen zu einem robusteren Gesamteindruck. Vielmehr reicht ein einzelner Restzweifel, um zur Aussage zu führen, dass nicht ausgeschlossen werden kann, dass diese Aussage auch ohne Erlebnisgrundlage zustande gekommen ist. Wie die Gutachter Fiedler und Schmid in ihrem Gutachten korrekt ausgeführt haben, hängt extrem viel von der Polung der ersten Fragestellung ab. Sucht man also nach dem Haar in der Suppe, will man also auf jeden Fall Falschpositive vermeiden, wird man eine Fehleinschätzungstendenz dahingehend haben, dass auch auf realem Erleben basierende Aussagen als möglicherweise nicht erlebnisbasiert bezeichnet werden. Das ist genau das, was derzeit in der Glaubhaftigkeitsbegutachtung passiert. Man erhebt eine Fülle von Beobachtungen und verwirft alles, was für den Wahrheitsgehalt und Realitätsgehalt der Aussage spricht, wenn nur eine Frage z. B. die Frage suggestiver Einflüsse offen bleibt. Einzelbeobachtungen mit der deduktiven Falsifikation werden nicht induktiv-statistisch zu einem Gesamtbild aggregiert. Also Stroh bleibt Stroh und wird nicht zu Gold und ein einzelner hypothetischer Restzweifel kann einen goldschimmernden Gesamteindruck zerstören, was die tatsächliche Irrtumswahrscheinlichkeit massiv erhöht.

Vor dem Hintergrund der zuvor ergangenen Polygrafenentscheidung, bei der auch den beiden hier wieder eingeladenen Sachverständigen Fiedler und Steller gefolgt worden war, kann man sich die Argumentationsnot des BGH vorstellen. Damals ging es tatsächlich um ein Untersuchungsverfahren und es lagen Forschungsergebnisse vor, die es erlaubten, Trefferwahrscheinlichkeit und Fehlerwahrscheinlichkeit zu quantifizieren. Nun, bei der Glaubhaftigkeitsbegutachtung ging es um eine Untersuchungsmethode, deren Trefferwahrscheinlichkeit deutlich geringer war, die aber trotzdem als adäquates Vorgehen offensichtlich erwünscht war und ja auch zugegebenermaßen im Vergleich zur Praxis vor dem Urteil wenigstens eine Standardisierung des Vorgehens mit sich brachte (vgl. König & Fegert, 2009). So konnte ein Verfahren, dessen Treffsicherheit in Bezug auf Lügen deutlich über dem Zufall liegt, grundsätzlich als völlig ungeeignetes Beweismittel im Verfahren nicht angewandt werden, da die Spezifität zu gering ist. Gleichzeitig konnte im hier dargelegten anderen Fall ein Vorgehen, dessen teilweise empirische Überprüfung zu lamentablen Ergebnissen geführt hat, als geeignet angesehen werden, weil es auf jeden Fall bei Vorliegen eines geringsten Restzweifels durch die durch nichts gerechtfertigte binäre Ergebnisdarstellung (wie bei einer Hypothesen-Überprüfung, hier aber ohne Überprüfungsregel) zu einem Ergebnis im Zweifel für die_den Angeklagte_n kommt.

„Se non è vero è molto ben trovato" – Wenn es nicht wahr ist, dann ist es doch sehr gut erfunden

Wie konnte es sein, dass diese von Missverständnissen und Widersprüchen belastete Entscheidung zu einem quasi unhinterfragten Dogma wurde? Zunächst einmal löst dieses Vorgehen viele praktische Probleme, weil eine zentrale Vorentscheidung bei der Beweiswürdigung aus den Verfahren auf die Expert_innen-Ebene verlagert werden kann. Dies erleichtert die Abwicklung solcher Verfahren. Instanzgerichte müssen sich an die Rechtsprechung und die Vorgaben des BGH halten und bislang haben Betroffene noch keine Anwält_innen gefunden, die mit diesem Vorgehen verbundene mögliche Verfassungsverstöße vor das Bundesverfassungsgericht gebracht haben. Aus kinder- und jugendpsychiatrischer Sicht bedeutet dies, dass im Einzelfall abgewogen werden muss, ob psychisch belasteten Kindern und Kindern mit Beeinträchtigungen der Teilhabe vor allem im sprachlich-kommunikativen Bereich die relativ aussichtslose Situation einer Strafanzeige oder eines Strafverfahrens überhaupt zugemutet werden sollte. Diese Frage erscheint auf jeden Fall berechtigter als die, trotz gegenlautender eindeutiger Klarstellung u. a. durch das Bundesministerium für Justiz, immer wieder von den Strafverfolgungsbehörden geäußerte Empfehlung bis zur strafrechtlichen Abwicklung auf eine Psychotherapie zu verzichten, um suggestive Einflüsse durch die Therapie auf die Aussage zu vermeiden. Der strafrechtlich wichtige und berechtigte Zweifelsgrundsatz „in dubio pro reo", der für die Tatbeurteilung durch das Tatgericht zentral handlungsleitend sein muss, wird durch die Methode, welche auf Restzweifel fokussiert, in das Gutachten vorverlagert. Damit wird die Beantwortung der Frage nach der Glaubhaftigkeit einer Aussage zu einer letztendlich binären Ja-Nein-Entscheidung. Eine differenzierte Darstellung von Plausibilitäten und Irrtumswahrscheinlichkeiten erfolgt in der Regel nicht. Viele Strafverfahren werden bei einer solchen Ausgangslage ohne Vorliegen anderer Befunde vor der Eröffnung einer Hauptverhandlung eingestellt. Soweit – so problematisch. Letzten Endes zählt nicht das Gesamtbildselbst bei einer ganz überwiegenden Wahrscheinlichkeit für den Realgehalt eines Tatvorwurfs bringen Restzweifel das ganze Gebäude zum Einsturz. Komplexe Plausibilitäten und Wahrscheinlichkeiten werden im vorgelagerten Gutachten auf die binäre Entscheidung reduziert. Dies widerspricht dem Wesen der richterlichen Überzeugungsbildung, bei der es weder absolute Gewissheiten noch den de facto a priori Ausschluss einzelner (zulässiger) Beweismittel durch eine isolierte Würdigung gibt. Scheinbar wissenschaftstheoretisch wird dieses Vorgehen wird mit dem Begriff des Ver-

werfens einer Hypothese legitimierend umschrieben. Tatsächlich testen wirklich hypothesenüberprüfende Verfahren eine Nullh##ypothese gegen eine Unterschiedshypothese oder Alternativhypothese und führen aufgrund einer Analyse von Stichproben im Gruppenvergleich zur Feststellung, ob ein signifikanter Unterschied vorliegt (Alternativhypothese) oder ob die „Nullhypothese" gilt (es liegt kein Unterschied vor). Hierzu werden statistische Signifikanztests verwendet. Diese Tests ermöglichen es auch, die Irrtumswahrscheinlichkeit für zwei Arten von Fehlern zu beschreiben. Nämlich Fehler erster Art (eine richtige „Nullhypothese" wird fälschlicherweise abgelehnt, auch Alphafehler genannt) und Fehler zweiter Art (eine richtige Alternativhypothese oder Unterschiedshypothese wird fälschlicherweise verworfen, Betafehler).

Überraschend ist, dass quasi nach Maslows „Law of the instrument" (Maslow, 1966) dieses Vorgehen auch in Kontexten angewandt wird, wo der Zweifelsgrundsatz nicht eine zentrale handlungsleitende Maxime ist, sondern wo die höchstrichterliche Rechtsprechung und jetzt auch das neue soziale Entschädigungsrecht im SGB XIV von *Plausibilitäten und Wahrscheinlichkeiten* bei der Tatfeststellung sprechen. Es resultiert eine epistemische Ungerechtigkeit (Fegert, 2022; Fricker, 2007; Fricker, 2023), ein systematisches Nicht-Hören des Wissens der Betroffenen, ein systematisches Zweifeln an der Glaubhaftigkeit ihrer Aussagen, sodass oft die rechtliche Abwicklung, in die die Betroffenen sehr viel Hoffnung setzen, den Schaden mehrt, den die Taten gesetzt haben.

Die rhetorischen Anleihen bei hypothesenüberprüfenden Verfahren erwecken offensichtlich wenigstens für viele Jurist_innen den Eindruck, dass das Vorgehen bei der Glaubhaftigkeitsbegutachtung in Deutschland der Anwendung eines wissenschaftlichen hypothesenüberprüfenden Verfahrens entspreche. De facto erfolgt aber keine Testung oder Hypothesenüberprüfung in diesem Sinne. Über Irrtumswahrscheinlichkeiten können keine Aussagen getroffen werden. Während also bei einem Test die Sensitivität, welche in Betracht zieht, wie viele Merkmale von einem bestimmten Test übersehen werden, von der Spezifität, bei der es darum geht, wie viele Merkmale fälschlicherweise auf das Vorliegen eines bestimmten Merkmals hindeuten, unterschieden werden kann, wird bei der Glaubhaftigkeitsbegutachtung durch eine Setzung, die sich vom Zweifelsgrundsatz her ableitet, ein extremer Akzent in Bezug auf die Spezifität gesetzt. In Bezug auf sexuellen Missbrauch im Kindesalter würde möglichst hohe Spezifität bedeuten, dass das Risiko falsch positiver Wertung möglichst vermieden wird. Bei diesem Vorgehen nimmt man hohe Verluste bei der Sensitivität in Kauf, d.h. man akzeptiert das Risiko mit diesem Vorgehen, real stattgefundenen Missbrauch, also einen tatsächliche Erlebnisbasierung fälschlicherweise zu verneinen. Bei der „Methode" der Glaubhaftigkeitsbegutachtung gibt es keine empirischen Angaben zu Sensitivität und Spezifität bzw. zu den Irrtumswahrscheinlichkeiten, sondern das Vorgehen arbeitet mit einer axiomatischen Setzung – orientiert am Zweifelsgrundsatz im Strafverfahren, die unterstellt, dass bei möglichen suggestiven Einflüssen tatsächlich Zweifel im Sinne des Grundsatzes „in dubio pro reo" nicht ausgeräumt werden können. Über die Wahrscheinlichkeit möglicher suggestiver Einflüsse wir dabei nichts gesagt. Im Gegensatz dazu, können wissenschaftliche Tests durch Methoden, wie z.B. einer Receiver-Operator-Charakteristik (ROC) das Verhältnis von Sensitivität und Spezifität quantifizieren und damit deutlich machen, ob bestimmte Verfahren eher als Screening-Verfahren (höhere Sensitivität mit höherer Irrtumswahrscheinlichkeit) oder eher zur sicheren Diagnostik geeignet sind. Ideal in diesem Sinne sind Testverfahren oder Untersuchungsverfahren, die eine maximal hohe Sensitivität und Spezifität vereinen. Dies gilt z.B. für die neueren AIDS-Tests, welche bei korrekter Anwendung quasi jeden Fall erkennen und keine Fehldiagnosen mehr liefern. Dieses Beispiel macht deutlich, wie weit die Methode der Glaubhaftigkeitsbegutachtung von einem wissenschaftlichen Test entfernt ist, zu dem empirische Aussagen über die Methode vorliegen, die es Anwender_innen oder Interpret_innen ermöglichen, Treffsicherheit und Irrtumswahrscheinlichkeit zu erläutern und bei der Interpretation der Ergebnisse zu berücksichtigen. Äußert man sich in einem Gutachten z.B. zur Intelligenz und hat mit einem standardisierten Intelligenztest einen bestimmten Wert als Ergebnis herausgefunden, kann man aufgrund der zugrunde liegenden Berechnungen auch den Range (Konfidenzintervall) angeben zwischen welchen Werten der wahre Wert mit 95%iger oder 99%iger Sicherheit liegt. Das heißt: Solche Methoden können die Wahrheit nicht 100%ig herausfinden, sie können aber die Wahrscheinlichkeit ebenso angeben wie Irrtumswahrscheinlichkeiten. Ganz anders die Glaubhaftigkeitsbegutachtung, die anhand von qualitativen Merkmalen, also Textinformationen, bei irgendeinem Restzweifel zur Bejahung der hier in irreführender Weise sogenannten „Nullhypothese" führt.

Fazit für die Wissenschaftstheorie in der KJP

Zwar wurden erste wissenschaftstheoretische Überlegungen von Philosophen wie Aristoteles schon in der Antike getroffen, zum Beispiel die Gegenüberstellung des induktiven und des deduktiven Schließens. Tatsächlich begann

die moderne Wissenschaftstheorie in den 20er Jahren des letzten Jahrhunderts mit dem Positivismus (logischer Empirismus). Gefordert wurde eine Verifizierbarkeit für die Überprüfung der Wahrheit einer Aussage eines physikalischen Regelsatzes, z. B. wurden Axiome der Physik für verifizierbar gehalten, während Metaphysisches nicht verifiziert werden könne. Induktionisten, wie z. B. Carnap, gingen in der Tradition des logischen Empirismus davon aus, dass induktiv auf Gesetzmäßigkeiten geschlossen werden kann: Durch die Ansammlung von Einzelbeobachtungen wäre es also möglich, grundsätzliche Regeln zu beschreiben. Karl Popper setzte dem das Prinzip der Falsifizierbarkeit entgegen, welches er philosophisch im Konzept des kritischen Realismus weiterentwickelte.

Gutachtende müssten die Stärken und Schwächen ihrer Methoden, ebenso wie die Prämissen der Anwendung, so verständlich erläutern, dass methodische Laien sie verstehen können. Der prinzipielle Gegensatz zwischen einem induktiven Vorgehen und dem daraus resultierenden Schluss auf die wahrscheinlich zugrundeliegende Wahrheit und einem deduktiven Vorgehen, welches eine Hypothese überprüft, sollte als wissenschaftstheoretisches Allgemeinwissen jedem bekannt sein. Die Rezeption von zwei Gutachten im Rahmen einer BGH-Entscheidung zur Glaubhaftigkeit macht aber deutlich, dass diese wissenschaftstheoretischen Prämissen nicht berücksichtigt und assoziativ Argumente zusammengeführt wurden, die einerseits den Eindruck erwecken sollten, dass eine deduktionistische Hypothesenüberprüfung erfolgte, bei der es nach bestimmten Regeln – statistisch z. B. beim Nichterreichen der festgesetzten Signifikanzschwelle – zum Verwerfen der Hypothese kommen muss. Und dass gleichzeitig durch ein induktives Vorgehen – welches ja von Deduktionisten abgelehnt wird – die Aussagekraft einzelner Merkmale durch eine statistische Aggregation gesteigert wird. Beide Elemente werden im Richterspruch des BGH zur Glaubhaftigkeitsbegutachtung vermischt. Woraus eine Scheinsicherheit in Bezug auf die Adäquanz des Vorgehens m suggeriert wird.

Die hier erfolgte wissenschaftstheoretische Erläuterung der Prämissen und die dezidierte Darstellung, dass es sich bei der angewandten Methode nicht um ein Testverfahren, ja noch nicht einmal um ein über die Anwendung des allgemeinen Grundsatz der induktiven statistischen Aggregation abgesichertes und in der Aussagekraft verstärktes Vorgehen handelt, sollte Rechtswissenschaftler_innen anregen, darüber nachzudenken, wie die Vorwegnahme einer Beweiswürdigung durch ein Gutachten, welches zu einer „Ja-Nein-Entscheidung" mit einer absolut ungleichen Risikogewichtung kommt, mit dem fundamentalen Grundsatz der freien richterlichen Beweiswürdigung in § 261 StPO vereinbar ist. Hier ging es nicht primär um diese für manche Betroffene mit zahlreichen bösen Überraschungen und negativen Konsequenzen verbundene Rechtspraxis, sondern in einem Themenheft zur Wissenschaftstheorie sollte deutlich gemacht werden, welche Auswirkungen die Nichtberücksichtigung wissenschaftstheoretischer Rahmenbedingungen hat. Wissenschaft ist gekennzeichnet dadurch, dass alles im Fluss ist. Die juristisch dogmatische Festlegung einer Quasi-Beweisregel hat im letzten Vierteljahrhundert den wissenschaftlichen Fortschritt im Bereich der Aussagepsychologie auf jeden Fall nicht gefördert. Dadurch, dass ein bestimmtes Vorgehen, welches mehr oder weniger einen Sonderweg in den deutschsprachigen Ländern darstellt, höchstrichterlich determiniert wurde, hat sich aber die Praxis vereinheitlicht, was auf jeden Fall schon eine gewisse Form der Qualitätssicherung darstellt, die hier nicht bestritten werden soll.

Gleichzeitig wurde aber sehr viel mehr Zeit in die Lehre und Schulung dieser Vorgehensweise investiert als in die weitere, selbst vom BGH angemahnte Erforschung ihrer unterschiedlichen Anwendungsbedingungen. Zum Verständnis wissenschaftlicher Methoden, wie zur Beweiswürdigung, gehört auch ein klares Wissen um deren Limitationen. Ziel dieses Beitrags war es, wissenschaftstheoretische Missverständnisse, aufgrund von ausführlichen Originalzitaten, offenzulegen, um insbesondere im Jahr eines möglichen Aufbruchs im Sozialen Entschädigungsrecht, in dem es tatsächlich darum geht, bei der Tatfeststellung den relativ am wahrscheinlichsten Tatablauf (§ 117 Abs.2 SGB XIV zu würdigen, vor der unkritischen Übernahme eines solchen, für die scheinbaren Bedürfnisse des Strafverfahrens, einseitig zuungunsten von Betroffenen verzerrten Vorgehens zu warnen. Hier scheint es unmittelbar geboten, das methodische Vorgehen in Bezug auf die Glaubhaftmachung von Taten im Kontext des SGB XIV von den Prämissen des BGH in Strafsachen zu lösen. Darüber hinaus wäre es endlich an der Zeit, auch durch empirische Forschung in Deutschland Spezifika insbesondere bei Aussagen von Kindern und Jugendlichen mit psychischen Störungen besser zu erforschen.

Literatur

Adorno, T.W. (Ed.). (1969). *Der Positivismusstreit in der deutschen Soziologie*. Berlin: Luchterhand.

Fegert, J. M. (2022). *Anerkennung psychischer Traumafolgen: Eine Spurensuche, inspiriert von der St. Michaelsfigur im Ulmer Münster*. Psychiatrie Verlag, Imprint BALANCE buch + medien verlag.

Fegert, J.M., Gerke, J. & Rassenhofer, M. (2018). Enormes professionelles Unverständnis gegenüber Traumatisierten: Ist die Glaubhaftigkeitsbegutachtung und ihre undifferenzierte Anwendung in unterschiedlichen Rechtsbereichen eine Zumutung für von sexueller Gewalt Betroffene? *Nervenheilkunde, 37*(07/08), 525–534. https://doi.org/10.1055/s-0038-1668320

Fegert, J.M., Gerke, J., Kliemann, A., Pusch, M., Rixen, S. & Sachser, C. (2024). *Expertise: Die Methode der forensischen Glaubhaftigkeitsbegutachtung im deutschen Sprachraum – Ein interdisziplinäres Plädoyer für eine kritische Bestandsaufnahme zur Anwendung der sogenannten „Nullhypothese" in unterschiedlichen Verfahrenskontexten*. Berlin: Arbeitsstab der Unabhängigen Beauftragten für Fragen des sexuellen Kindesmissbrauchs. Verfügbar unter: https://beauftragte-missbrauch.de/fileadmin/user_upload/Materialien/Publikationen/Expertisen_und_Studien/Expertise_Glaubhaftigkeitsbegutachtung.pdf

Fiedler, K. & Schmid, J. (1999). Gutachten über Methodik und Bewertungskriterien für psychologische Glaubwürdigkeitsgutachten. *Praxis der Rechtspsychologie, 9*(2), 5–45.

Fricker, M. (2007). *Epistemic Injustice. Power and die Ethics of Knowing*. Oxford University Press.

Fricker, M. (2023). *Epistemische Ungerechtigkeit. Macht und Ethik des Wissens*. München: C. H. Beck.

Laney, C. & Takarangi, M.K. (2013). False memories for aggressive acts. *Acta psychologica, 143*(2), 227–234. https://doi.org/10.1016/j.actpsy.2013.04.001

Maslow, A.H. (1966). *The psychology of science: A Reconnaissance*.

Medizinischer Fakultätentag. (o.D.). *Positionspapier: Vermittlung von Wissenschaftskompetenz im Medizinstudium*. Abgerufen am 16.5.2024 unter https://medizinische-fakultaeten.de/wp-content/uploads/2018/01/Positionspapier-Wissenschaftlichkeit.pdf

Niehaus, S. & Krause, A. (2023a). Threats to scientifically based standards in sex offense proceedings: Progress and the interests of alleged victims in jeopardy. *Monatsschrift für Kriminologie und Strafrechtsreform, 106*(3), 165–183. https://doi.org/10.1515/mks-2023-0018

Niehaus, S. & Krause, A. (2023b). Wissenschaftsorientierung in Sexualstrafverfahren in Gefahr: Fortschritte und Opferinteressen stehen auf dem Spiel. *Praxis der Rechtspsychologie*, *33*(2). https://irf.fhnw.ch/handle/11654/38002

König, C. & Fegert, J.M. (2009). Zur Praxis der Glaubhaftigkeitsbegutachtung unter Einfluss des BGH-Urteils. *Interdisziplinäre Zeitschrift der Deutschen Gesellschaft gegen Kindesmisshandlung und -vernachlässigung e.V., 12*(2), 16–41.

Sauerland, S. & Waffenschmidt, S. (2018). Welche Halbwertszeit hat medizinisches Wissen. *KVH Journal, 6*, 20–22.

Shaw, J. (2020). Do false memories look real? Evidence that people struggle to identify rich false memories of committing crime and other emotional events. *Frontiers of Psychology, 11*(10), 3389. https://doi.org/10.3389/fpsyg.2020.00650

Steller, M. & Köhnken, G. (1989). Criteria-based statement analysis. In D.C. Raskin (Ed.), *Psychological methods in criminal investigation and evidence* (pp. 217–245). Springer Publishing Company.

Steller, M. & Volbert, R. (1999). Forensisch-aussagepsychologische Begutachtung (Glaubwürdigkeitsbegutachtung). Wissenschaftliches Gutachten für den Bundesgerichtshof. *Praxis der Rechtspsychologie, 9*(2), 46–112.

Volbert, R. (1995). Glaubwürdigkeitsbegutachtung bei Verdacht auf sexuellen Mißbrauch von Kindern. *Zeitschrift für Kinder- und Jugendpsychiatrie, 23,* 20–26.

Historie
Manuskript eingereicht: 21.05.2024
Manuskript akzeptiert: 04.09.2024
Onlineveröffentlichung: 16.10.2024

Interessenkonflikte
Es bestehen keine Interessenkonflikte.

Förderung
Open-Access-Veröffentlichung ermöglicht durch die Universität Ulm.

ORCID
Jörg M. Fegert
https://orcid.org/0000-0001-6070-4323
Cedric Sachser
https://orcid.org/0000-0002-9353-7936
Jelena Gerke
https://orcid.org/0000-0002-6338-7404

Prof. Dr. med. Jörg M. Fegert
Klinik für Kinder- und Jugendpsychiatrie/Psychotherapie
Universitätsklinikum Ulm
Steinhövelstr. 5
89075 Ulm
Deutschland

joerg.fegert@uniklinik-ulm.de