U.Mortensen

Chicken Sexing, Diagnose, und Prof. Gigerenzers Portefolio

Den Hinweis auf Gigerenzers Intuition-Theorie (Dank an den Kommilitonen) möchte ich als Anlaß nehmen, das Thema noch einmal etwas genauer anzusprechen. Der Kürze wegen (man kann es von der Web-Seite aus anklicken) informieren, obwohl das Interview das Thema doch ein wenig verkürzt widerspiegelt.

http://www.zeit.de/campus/2007/03/interview-gigerenzer

Gigerenzers Thesen widersprechen ja nicht den Befunden über intuitives Urteilen, die ich in der Vorlesung genannt habe, denn er weist ja darauf hin, dass bei diesen Fähigkeiten Training notwendig sei. In diesem Zusammenhang ist auch nicht erst seit 2007 (Zeitpunkt des Interviews) bekannt, wie sicher gewisse trainierte Fertigkeiten gewissermaßen ohne Nachdenken durchgeführt werden können, und was Gigerenzer über sein Portfolio schreibt, läßt sich vermutlich auch ohne Rückgriff auf die Intuition erklären, sondern durch die Hintergründe für die Bekanntheit bestimmter Firmen auch bei Menschen, die nicht primär mit Wirtschaft zu tun haben und die diese Hintergründe auch gar nicht explizit kennen (wer Siemens- und VW-Aktien kauft, hat gute Chancen, auch wenn er nichts über diese Firmen weiß, etc).

Zum Training gehört, dass Feedback gegeben wird. Bei diagnostischen Übungen von "Experten" fehlt häufig nicht die Erfahrung, sondern das Feedback. Die folgenden Abschnitte sind dem Text *Verstehen oder Erklären? Die Rolle experimenteller und statistischer Methoden in der modernen Psychologie* entnommen, das auf meiner Webseite unter Texte gefunden werden kann. In den folgenden Abschnitten geht es hauptsächlich um die Tversky-Kahnemann-Arbeiten. Weiter möchte ich auf den Artikel von Horsey aufmerksam machen, den Sie anklicken können.

1 Zur Intuition in der Diagnostik

1.1 Verstehende versus "mechanische" Diagnostik

Wie in Deutschland, hat es auch in den USA seit den 70-er Jahren einen großen Anstieg der Positionen für beratend, gutachtend und therapeutisch tätigen Psychologen gegeben. Die Ausbildung ist an vielen Universitäten berufsbezogen gewesen, d.h. es wurde großes Gewicht auf die Einübung praktischer Fertigkeiten, aber geringes Gewicht auf die Einübung wissenschaftlichen Denkens gelegt. Die grundlegende Annahme war (und ist es zum Teil immer noch), dass man durch Erfahrung zum Experten wird. Psychologen werden dieser Auffassung nach durch hinreichend lange Tätigkeit zum Experten, und sie berufen sich auf ihre Erfahrung, wenn sie Diagnosen stellen und Voraussagen über zukünftiges Verhalten (z.B. bei Strafgefangenen) machen. Ebenso berufen sich Therapeuten auf ihre Erfahrung,

wenn sie die Ursachen für psychische Störungen diagnostizieren und eine dementsprechende Therapie vorschlagen. Ein zentrales Argument ist dabei, dass diese Erfahrung nicht durch "mechanisches" Urteilen etwa anhand von Tests und Fragebögen ersetzt werden könne. Überdies sei das Fällen von Entscheidungen (z.B. über die Freilassung eines Gewalttäters nach Verbüßung eines Teils oder seiner gesamten Haftstrafe) ein "dehumanisierender Akt", da sich Menschen nun mal nicht in Zahlen fassen ließen.

Es hat schon früh Versuche gegeben, solche Behauptungen zu überprüfen. Dazu werde noch einmal zusammengefaßt, worum es hier geht. Postuliert wird, dass es zwei Arten der Diagnosen gibt:

- die Diagnose auf der Basis des Verstehens: sie basiert auf professionellem Training, Erfahrung, Einsicht in die individuelle Persönlichkeitsstruktur, und
- Die Diagnose und Vorhersage anhand statistischer Formeln ("mechanische" oder "aktuariale" (englisch: "actuarial") Diagnose)[1].

Der Verstehende Ansatz dürfte nunmehr klar sein. Der "mechanische" Ansatz funktioniert wie folgt: Es sei Y eine Maßzahl für das vorherzusagende bzw. zu diagnostizierende Merkmal, und $X_1, X_2, \ldots, X_n$ seien Maßzahlen für die Merkmale, anhand derer die Vorhersage getroffen werden soll. Die Vorhersage Y wird dann gemäß

$$Y = b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \tag{1}$$

berechnet. Dabei sind die Zahlen $b_1, \ldots, b_n$ sogenannte "Gewichte" , mit denen die durch die X_j, $j = 1, 2, \ldots, n$ repräsentierten Symptome bei der Vorhersage von Y gewichtet werden. Sie sind für das vorauszusagende Merkmal spezifisch und werden aus den Daten von Voruntersuchungen geschätzt, wobei die Schätzung bestimmten Optimalitätskriterien genügt. Die Diagnose oder Vorhersage wird (abfällig) "mechanisch" genannt, weil sie, sind die X_j erst bekannt, gewissermaßen durch bloßes Ausrechnen erfolgt: die Entscheidungen werden nach Maßgabe des Wertes von Y getroffen. Die Kritik der "Verstehenden" Psychologen und Experten richtet sich gegen die "rigide" Fokussierung auf einen festen Satz von Prädiktoren oder Symptomen X_j und deren ebenso "rigide" Gewichtung; die lebensweltliche Realität und Historizität des lebendigen Individums sei nun einmal nicht in derart einfache Formeln zu pressen sondern müsse ganzheitlich erfahren werden.

Die Preisfrage ist, welcher der beiden Ansätze tatsächlich besser ist.

Meehl (1954) präsentierte eine erste Zusammenschau verschiedener Arbeiten zu dieser Frage. Er analysierte 20 Studien, in denen (i) Experten verschiedene Größen vorhergesagt hatten, und in denen (ii) die gleiche Vorhersage an den gleichen Individuen anhand "mechanischer" Tests bzw. Fragebögen gemacht wurden. Die vorzusagenden Variablen waren z.B. akademischer Erfolg (vorhergesagt bei Studienanfängern), die Reaktion auf eine Elektroschocktherapie, die Wahrscheinlichkeit

[1] Der Ausdruck "actuarial" bezieht sich metaphorisch auf Formeln, die Lebensversicherungen für die zu erwartende Lebenszeit eines Versicherungsnehmers benutzen. Diese Formeln haben einen analogen Aufbau wie die in der "mechanischen" Diagnostik verwendeten. Im Deutschen ist dieser Ausdruck weniger gebräuchlich.

eines Rückfalls bei Strafgefangenen, etc. Die "mechanische" Vorhersage beruhte für den akademischen Erfolg auf einer Gewichtung von Schulnoten und Resultaten in Fähigkeitstests, für die Elektroshocktherapie auf einer Gewichtung von Ehestand, Dauer der psychotischen Störung, Schätzung ("Rating" auf einer Skala) der "Einsicht" des Patienten in seinen Zustand, bei den Strafgefangenen auf einer Gewichtung der Anzahl der vorangegangenen Verurteilungen sowie der Beurteilung im Gefängnis. Es ergab sich das

> **Resultat:** Die Vorhersage der "verstehenden" und "erfahrenen" Experten war in keinem Fall besser als die "mechanische" Vorhersage.

Die Vorhersage anhand gewichteter Prädiktoren war also ebenfalls nicht perfekt, aber sie war *besser*, d.h. hatte eine geringere Fehlerquote, als die Vorhersagen der *Verstehenden* Psychologen oder Experten. Die Wahl anderer Symptome der Prädiktoren hätte die mechanische Vorhersage womöglich noch verbessert, grundsätzlich muß aber gesagt werden, dass eine perfekte Vorhersage von Verhalten wohl aus prinzipiellen Gründen nicht möglich ist. Von ebenso grundsätzlicher Bedeutung ist aber, dass die "mechanische" Diagnose stets besser ist als die auf angebliche "Einsicht" in die psychische Struktur des Beurteilten beruhende Diagnose.

Sawyer (1966) wiederholte diese Art von Untersuchung auf der Basis von 45 Studien, - das Ergebnis entsprach dem Meehls.

Smith & Glass (1977) analysierten weitere 350 Studien zur Frage verstehende versus mechanische Diagnose: das Ergebnis entspricht dem der bereits genannten Studien. Eine weiter führende Diskussion dieser Frage findet man in Dawes (1996).

Dawes berichtet ebenfalls über Studien zum Erfolg von Psychotherapie. Er faßt die Vielzahl der Studien zu diesem Thema wie folgt zusammen: Psychotherapie sei i.a. erfolgreich,

- aber unabhängig vom akademischen Grad des Therapeuten,
- und unabhängig von
  - der Art der applizierten Therapie,
  - mit leichtem Vorsprung für die Verhaltenstherapie, *sofern wohldefinierte Verhaltensprobleme behandelt wurden*, und
  - von der "Erfahrung" der/des Therapeutin/en.

Diese Resultate sind überraschend und gegenintuitiv: man sollte meinen, dass mit der Erfahrung auch die Kompetenz wüchse. Dawes führt zur Erklärung des vorstehenden Sachverhalts einige Ergebnisse von Lernexperimenten an. Demnach lernt man nur, wenn man geeignete Rückmeldungen ("Feedback") über die eigenen Handlungen bekommnt. Ob eine Therapie erfolgreich ist, ist aber oft über längere Zeitspannen hinweg nicht zu beurteilen. Hinzu kommen spezifische Gedächtnisfallen, in die auch ein guter Therapeut läuft: man merkt sich eher Erfolge und vergißt oder verdrängt Mißerfolge. Geht es einem Patienten in einer Stunde gut, so führt man dies auf die "richtige" Verhaltensweise (Diagnose, Empathie, etc) in

der letzten Stunde zurück und sieht darin also eine Bestätigung der Vorgehensweise. Geht es dem Patienten eher schlecht, so wird es dafür andere Gründe als die letzte Stunde geben. Vielfach wird ein (scheinbarer) Erfolg der Therapie ganz einfach durch den sogenannten Regressionseffekt erzeugt. Hierbei handelt es sich um einen rein statistischen, von der Behandlung unabhängigen Effekt; er soll kurz beleuchtet werden.

Zwischenbetrachtung: Der Regressionseffekt Der Regressionseffekt wird in der Statistik ausführlich diskutiert und soll hier deshalb nur in seiner einfachsten Form betrachtet werden. Im Kern bezeichnet der Ausdruck der Tatsache, dass eine zufällig variierende Größe mit größerer Wahrscheinlichkeit Werte in einer bestimmten Nachbarschaft ihres Mittelwertes annimmt als extreme Werte[2]. Registriert man von einander unabhängige Beobachtungen dieser Variablen und findet man bei einer bestimmten Beobachtung einen relativ großen oder kleinen Wert ("groß" und "klein" sind dabei relativ zur Streuung der Werte definiert), so wird man mit großer Wahrscheinlichkeit bei der jeweils nachfolgenden Beobachtung einen kleineren oder größeren, also näher am Mittelwert liegenden Wert finden. Die folgenden Betrachtungen basieren auf einer starken Vereinfachung und dienen nur der Illustration des Prinzips.

Es werde einmal der Spezialfall angenommen, dass die Therapie überhaupt keinen Effekt hat, d.h. das Verhalten oder die Befindlichkeit des Patienten mögen nicht von den therapeutischen Sitzungen abhängen. Aufgrund einer Vielzahl von Einflüssen schwankt die Befindlichkeit des Patienten um einen gewissen Mittelwert herum, wobei i.a. extreme Werte aufgund allgemeiner statistischer Gesetzmäßigkeiten seltener vorkommen als solche, die in der Nachbarschaft des Mittelwertes liegen. Ist sie schlechter als ein bestimmter, kritischer Wert, greift der Therapeut mit einer bestimmten Behandlung ein. Mit großer Wahrscheinlichkeit, d.h. in den meisten Fällen, liegt nun unabhängig davon in der nächsten Sitzung die Befindlichkeit wieder näher am Mittelwert, also oberhalb des kritischen Wertes. Der Therapeut sieht dies als Erfolg seiner Behandlung und fügt dies seinem "Erfahrungsschatz" bei. Tatsächlich spiegelt die Beobachtung nur den Regressionseffekt wieder, d.h. den Trend von Beobachtungen, nach extremen Werten mit größerer Wahrscheinlichkeit wieder näher am Mittelwert zu liegen. Dieser Effekt ist auch wirksam, wenn es tatsächlich einen positiven Therapieeffekt gibt, wenn aber außerdem noch zufällige (d.h. nicht vom Therapeuten kontrollierbare) Effekte die Befindlichkeit des Patienten beeinflussen, - und die gibt es natürlich stets. Leider entzieht sich dieser Effekt einem nur 'wesentlich schauenden' (um einen Ausdruck der Diltheyschen Philosophie zu gebrauchen) und 'intuitiv nacherlebenden' Therapeuten ohne Kenntnis elementarer statistischer Sachverhalte.

Die Ergebnisse der Untersuchungen zur Frage, ob Vorhersagen auf der Basis eines "Verstehenden" oder eines "mechanischen" Ansatzes besser seien, lassen vermuten, dass die ersteren Urteile zumindest eine größere Fehlerkomponente haben als die mechanischen Urteile. Dies führt zu der Frage, inwieweit überhaupt von Therapeuten und Experten "verstanden" wird.

[2]Dies gilt nicht stets: bei bimodalen Verteilungen kann es sein, dass die Werte gerade nicht mit größerer Wahrscheinlichkeit in der Nachbarschaft des Mittelwertes liegen.

Goldberg (1960) legte eine Studie vor, in der die Klassifikation von Patienten als entweder "neurotisch" oder "psychotisch" von Experten gefordert wurde. Konkurrierend dazu wurde die Klassifikation anhand einer Formel, also "mechanisch" vorgenommen. Ein Patient galt als "neurotisch", wenn sie oder er an internem emotionalen Stress leidet, aber Kontakt zur externen Realität hat, und als "psychotisch", wenn zusätzlich wie bei der Schizophrenie der Kontakt zur externen Realität verloren gegangen ist. Alle Patienten sind mit dem Minnesota Multiphasic Personality Inventory (MMPI) getestet worden. Der MMPI besteht aus 567 Fragen ("Items"), die entweder mit "ja" oder mit "nein" zu beantworten sind. Beispiele für Items sind

- Manchmal bin ich zu nichts gut
- Mein Sex-Leben ist in Ordnung
- Ich mag Zeitschriften über Modellbau
- Ich finde es richtig, dass auf die Einhaltung des Gesetzes geachtet wird, etc

Die Antworten eines Probanden werden zu einem "Profil" von 10 "Scores" zusammengefasst; jeder dieser Scores ist indikativ für das Vorhandensein oder die Abwesenheit bestimmter psychischer Störungen. Die Aufgabe der Experten bestand darin, diese Profile zu interpretieren. Bei dieser Interpretation handele es sich um eine "Kunst", wie die Experten versichern, die aber von Experten gelehrt werde. Goldberg ließ mehr als 1000 (Tausend) solcher Profile interpretieren. Für jedes Profil wurde außerdem eine entsprechende "mechanische" Klassifikation vorgenommen; diese bestand im Wesentlichen in der Anwendung der Gleichung

$$Y_j = b_0 + b_1 X_{j1} + \cdots + b_p X_{jp} + e_j, \tag{2}$$

nachdem optimale Gewichte b_j gefunden worden waren.

Die "mechanische" Klassifikation lieferte in ca 70% der Fälle eine korrekte Diagnose. Von den Experten erreichte keiner diese Erfolgsquote. Eine unmittelbare Konsequenz dieses Befundes ist, dass Personen selbst dann, wenn sie auf eine von ihnen selbst als substantiell eingestufte Erfahrung zurückgreifen können, die gegebene Information offenbar nicht mit einer durch die Formel (**??**) gegebenen Optiomalität nutzen können. Dieser Sachverhalt wird weiter unten noch diskutiert werden.

Bloom & Brundage (1947) verglichen die Voraussagen von Experten, die insgesamt 37 000 Rekruten der US-Army interviewt hatten, bezüglich der zu erwartenden (militärischen) Karriere der Rekruten. Simultan dazu wurden bei dem gleichen Rekruten Leistungstests (Intelligenz- und ähnliche Tests) durchgeführt. Die "Scores", d.h. Punktwerte, in diesen Tests wurden zusammen mit den Schulnoten der Rekruten gewichtet und dann gemäß dem Ansatz (2) ebenfalls zu einer "mechanischen" Vorhersage der Karriere benutzt. Die Voraussagen der Experten *waren konsistent schlechter* als diese mechanische Voraussage. Es hat mehrere Replikationen dieser Studie gegeben, und alle kamen zu dem gleichen Ergebnis: Experten, die aufgrund ihrer "Erfahrung" und angeblich "ganzheitlicher" Einsicht urteilen, sind konsistent

schlechter als eine Beurteilung aufgrund optimal gewichteter Indikatoren (die x_i) gemäß (2) (vergl. Dawes, 1996).

Bekanntlich müssen sich Interessenten für ein Medizinstudium in den USA einem Eignungstest unterziehen. Dawes (1996) referriert hierzu die *Texas Medical Study (1979)*. So wurden - bis zum Jahr 1979 - jährlich 150 Studenten zur Texas Medical School in Houston zugelassen. Von ca 2200 Bewerbern wurden die am besten Qualifizierten 800 nach Maßgabe ihres Schulabschlusses etc ausgesucht. Jeder dieser 800 wurde anschließend nach Houston eingeladen und dort von je einem Mitglied des Zulassungsausschusses und einem Mitglied der medizinischen Fakultät interviewt. Die Interviewer reichten dann schriftliche Bewertungen der Bewerber bei einem zentralen Kommitee ein. Die Bewertung wurde von jedem Mitglied des Kommitees auf einer Skala von 0 (nicht akzeptabel) bis 7 (exzellent) beurteilt. Aufgrund dieses "Rankings" wurde dann eine Rangordnung der 800 in die engere Wahl gekommenen Bewerber aufgestellt. Die 150 schließlich an der Medical School aufgenommenen Studierenden gehörten alle zur Gruppe der 350 Studierenden an, die in den Interviews am besten abgeschnitten hatten.

1979 entschied der Staat Texas, dass von nun an nicht nur 150, sondern 200 Bewerber von der Medical School aufgenommen werden sollten. Es mußten also 50 weitere Bewerber aufgenommen werden. Diese sollten aus den 800 Bewerbern ausgewählt werden, die bis dahin nicht berücksichtigt, d.h. "herausgeprüft" worden waren. Von diesen waren aber nur noch diejenigen Bewerber verfügbar, die einen Rangplatz zwischen 700 und 800 bekommen hatten, d.h. nur noch die angeblich Schlechtesten der ursprünglich 800 Bewerber waren noch verfügbar, die besseren hatten sich inzwischen anderweitig orientiert. 86% dieser noch übrig gebliebenen Bewerber hatten weder innnerhalb von Texas noch außerhalb einen Studienplatz bekommen. Von diesen übrig gebliebenen Bewerbern wurden nun 50 ausgewählt, um auf die geforderte Anzahl von 200 Studienanfängern zu kommen.

Keinem der Professoren an der Medical School wurde mitgeteilt, welche Studierenden zu der Gruppe der "besten", zuerst ausgewählten 150 gehörten und welche zu der Gruppe der zuletzt ausgewählten 50 Bewerbern gehörten.

DeVaul et al. (1986) verglichen den Studienerfolg dieser 50 "Schlechtesten" mit dem der zuerst ausgewählten 150 "Besten". Das für die "Experten" ebenso verblüffende wie peinliche Resultat war, dass am Ende des zweiten Studienjahres keinen Unterschied hinsichtlich der Studienleistung zwischen diesen beiden Gruppen gab, ebenso wenig gab es einen Unterschied nach dem Ende der klinischen Ausbildung im vierten Jahr. Aus *jeder* dieser beiden Gruppen machten 82% den Dr. med. (M.D.), die besten Noten verteilten sich gleichmäßig auf beide Gruppen, etc. Bemerkenswert an diesem Resultat ist, dass die Bewerber mit den Rangplätzen 700 bis 800 auch an keiner anderen Universität zugelassen worden waren; - trotzdem waren sie im Durchschnitt genau so gut wie die 150 "Besten". DeVaul et al. schlossen aus diesen Ergebnissen, dass die Interviews (auf der Basis der Interviews wurden ja die Bewerber ursprünglich zugelassen oder auch nicht) eine *komplette Zeitverschwendung* waren.

Milstein et al. (1981) verglichen den Studienerfolg von Bewerbern an der Yale Medical School, die aufgrund von Interviews entweder zugelassen wurden oder

nicht. Nicht alle der Zugelassenen studierten dann tatsächlich in Yale, einige studierten an anderen Universitäten. Diese wurden nun mit denjenigen Studierenden verglichen, die ebenfalls an diesen anderen Universitäten studierten, aber von Yale abgewiesen worden waren. *Es konnten keine Unterschiede zwischen diesen beiden Gruppen gefunden werden*! Man kann also nicht sagen, dass die von Yale nicht zugelassenen Bewerber milder beurteilt wurden, weil sie nicht an einer Eliteuniversität wie Yale studierten und das *deshalb* kein Unterschied zu den von Yale Zugelassenen existierte: die Vergleichsgruppe hatte ja an den gleichen Universitäten, also nicht in Yale, studiert. Die Schlußfolgerung ist die gleiche wie die, die aus der Studie der Texas Medical School gezogen wurde: die Interviews durch "Experten" haben keine "Validität", d.h. sie können die wirklich relevanten Merkmale nicht erfassen und sind deshalb in der Tat reine Zeitverschwendung. Dawes (1996) führt aus, dass Interviewer offenbar auf ein ganz anderes Merkmal bzw. ein andere Fähigkeit als die Fähigkeit zu einem erfolgreichen Medizinstudium reagieren: es ist die Fähigkeit, beim Interviewer den Eindruck zu erzeugen, für das Studium geeignet zu sein. Das ist natürlich nicht dasselbe wie die tatsächliche Fähigkeit zu einem erfolgreichen Studium.

Ein ähnliches Bild ergibt sich, wenn Strafgefangene (insbesondere Gewalttäter) in bezug auf eine mögliche vorzeitige Entlassung interviewt und geprüft werden: die beste (d.h. die mit größter Wahrscheinlichkeit richtige) Voraussage ist, dass gewalttätiges Verhalten *nicht* wiederholt wird (Monahan, J., 1984). Aber nicht nur Laien, sondern auch Experten haben eine Neigung ("Bias"), zu glauben, dass eine Wiederholung von Gewalttaten normal sei. Die beste (wenn auch nicht sichere!) Vorhersage kann mithilfe gewichteter Symptome aufgrund der Gleichung (**??**), also "mechanisch", getroffen werden.

Die mangelnde Qualität von Beurteilungen durch Experten auf einer "verstehenden" Basis ergibt sich u.a. aus der Schwierigkeit, Informationen über zu beurteilende Personen in der richtigen Art und Weise zu verarbeiten; die Informationen werden nicht korrekt "gewichtet", zum Teil werden sie auch einfach übersehen. Deshalb sind nicht nur psychologische, sondern z.B. auch medizinische Diagnosen und Voraussagen oft mit systematischen Fehlern behaftet.

So untersuchte Einhorn (1972) die Einschätzung der Überlebensrate von Hodgkin-Patienten durch medizinische Experten für diese Krankheit. Die Experten schätzten die verbleibende Lebenszeit von insgesamt 193 Patienten. Unabhängig davon wurde die Überlebenszeit "mechanisch" durch optimale Gewichtung gemäß des Ansatzes (**??**) von 9 beobachtbaren Merkmalen der Patienten geschätzt.

Alle Patienten starben, so dass die tatsächliche Überlebenszeit mit der vorausgesagten verglichen weden konnte. *Während die Formel die Überlebenszeit relativ exakt vorhersagte, gab es keinerlei Zusammenhang zwischen den Vorhersagen der Experten (Ärzte) und den tatsächlichen Überlebensraten.*

Dawes (1996) berichtet, dass er diesen Fall in einem Vortrag vor Ärzten einer bekannten amerikanischen Medical School erwähnt habe. Die Reaktion der Ärzte auf diesen Befund bestand darin, die mangelnde Vorhersagefähigkeit der Experten durch deren mangelnde Qualität zu erklären; in dieser Reaktion stimmen sie mit anderen Experten überein, deren Urteile sich nicht als valide erweisen. Der Dekan

der Medical School verwies u.a. auf einen bekannten Experten, den berühmten Dr. XY, der bekannt für seine exakten Vorhersagen sei. Diese Argumentation sei, so Dawes, typisch für Personen, die sich als Experten fühlen: empirische Resultate zur Fehlerhaftigkeit von Expertenurteilen werden durch die mangelnde Qualität der in der Untersuchung evaluierten Experten erklärt. Die Möglichkeit, dass diese Art von Befunden aus der grundsätzlichen Beschränkung der Informationsverarbeitungsfähigkeiten von Menschen resultiert, wird - irrationalerweise - nicht akzeptiert. Denn auch der berühmte Dr. XY war einer der von Einhorn evaluierten Experten, und die Korrelation zwischen seinen Vorhersagen und den tatsächlichen Überlebenszeiten der Hodgkin-Patienten[3] war gleich Null! Bei Beurteilungen wird von beobachtbaren Eigenschaften, etwa eines Menschen, auf das Vorhandensein oder Nichtvorhandensein anderer Merkmale geschlossen. Repräsentiert A die Menge der beobachteten Eigenschaften und B die Menge der erschlossenen Eigenschaften, so beruht eine Beurteilung i.a. auf der Annahme einer Beziehung der Art "Wenn A, dann auch B". Bei einem komplexen System wie dem Menschen gelten solche Regeln i.a. nicht deterministisch, sondern stochastisch[4], Regel d.h. sie sind von der Art "Wenn das Merkmal A vorliegt, dann liegt mit der Wahrscheinlichkeit $P(B|A)$ auch das Merkmal B vor". $p(B|A)$ bezeichnet hier die *bedingte* Wahrscheinlichkeit von B, gegeben A, bezeichnet. Im allgemeinen ist die *unbedingte* Wahrscheinlichkeit von B nicht gleich der bedingten Wahrscheinlichkeit $p(B|A)$: hat man keinerlei Information über das morgige Wetter, so ist die Wahrscheinlichkeit $p(Regen)$, $B = Regen$, dass es morgen regnen wird, anders einzuschätzen als im Falle der Kenntnis der Wettervorhersage (A). Für $P(B|A) \approx 1$ erhält man den Spezialfall einer deterministischen Vorhersage. Die Tatsache, dass man nur stochastische Vorhersagen über Verhalten, Erkrankungen etc treffen kann, ergibt sich daraus, dass ein Verhalten B nicht nur eine mögliche Ursache A haben kann. Der Umgang insbesondere mit bedingten Wahrscheinlichkeiten ist, wenn man keinen Taschenrechner zur Hand hat, nicht ganz einfach; Menschen neigen dazu, $P(B|A) \approx p(A|B)$ anzunehmen. Welche Implikationen diese Annahme haben kann, wenn sie nicht gerechtfertigt ist, wird im folgenden Abschnitt erläutert.

1.2 Asymmetrien bei Vorhersagen

Grundsätzlich ergeben sich Asymmetrien der Vorhersage aus der Tatsache, dass aus der Aussage "Wenn A, dann auch B" die Aussage "Wenn B, dann auch A" *nicht* folgt. Das folgende Beispiel erläutert diesen Sachverhalt.

Beispiel 1.1 Aus den Untersuchungen des (Brust-) Gewebes von an Brustkrebs erkrankten Frauen hat sich ergeben, dass viele dieser Frauen ein "Risikogewebe" RG, d.h. ein vom Gewebe nicht an Brustkrebs erkrankter Frauen in bestimmter Weise abweichendes Gewebe haben; die Details hierzu sind für das Folgende nicht

[3] Eine Korrelation von Null bedeutet, dass es keinen systematischen Zusammenhang zwischen der tatsächlichen Ausprägung des zu beurteilenden Merkmals und dem Urteil der Experten über diese Ausprägung gibt. Dies bedeutet nicht, dass es keine korrekten Vorhersagen der Merkmalsausprägung gibt, denn bei rein zufällig getroffenen Vorhersagen kommt es ja auch zu korrekten Vorhersagen.

[4] stochastisch: den Zufall einbeziehend

von Belang. 57 % der Frauen haben ein Gewebe vom Typ RG, 43% haben es nicht. Bei einer Stichprobe von an Krebs erkrankten Frauen zeigte sich nun, dass 92% der *erkrankten* Frauen ein Gewebe vom Typ RG hatte. Daraus wurde geschlossen, dass die Wahrscheinlichkeit, an Brustkrebs zu erkranken, ca 90 % sei, wenn eine Frau das Gewebe vom Typ RG hat. Dementsprechend wurde Frauen, deren Brustgewebe vom Typ RG war, die aber (noch) nicht an Brustkrebs litten, empfohlen, eine *prophylaktische Mastektomie*, d.h. eine vorbeugende Entfernung des Brustgewebes vornehmen zu lassen (in einem Ort in Kalifornien folgten 90 Frauen im kritischen Alter von 40 bis 49 Jahren innerhalb von zwei Jahren dieser Empfehlung, vergl. McGee (1979))[5].

Weitere Untersuchungen ergaben, dass 7.7 % der Frauen ohne den Gewebetyp *RG* im Alter zwischen 40 und 59 Jahren an Brustkrebs erkranken, und 12 % der Frauen in diesem Altersabschnitt erkranken, wenn sie diesen Gewebetyp haben. Die Erkrankungswahrscheinlichkeit im Fall RG ist also tatsächlich 1.6-mal höher als im Falle keines Risikogewebes, *aber sie beträgt nicht 92 %).* □

Die beobachtete Asymmetrie der beobachteten Häufigkeiten liegt weder ein Rechen- noch ein Beobachtungsfehler zugrunde, sondern ergibt sich aus den entsprechenden bedingten Wahrscheinlichkeiten. Dies soll kurz erläutert werden.

Es sei $n(\text{Krebs})$ die Anzahl der an Krebs leidenden Frauen, $n(\text{RG})$ die Anzahl der Frauen mit Risikogewebe, etc. Die Gesamtzahl N einer Stichprobe von Frauen läßt sich einmal aufteilen in die Teilmenge derjenigen, die an Krebs leiden, und die Teilmenge denjenigen, die dies nicht tun, d.h. es gilt sicherlich

$$N = n(\text{Krebs}) + n(\text{kein Krebs}).$$

Zum anderen läßt sich die Stichprobe aufteilen in die Menge derjenigen Frauen, die das Risikogewebe haben, und die Menge der Frauen, die es nicht haben, so daß ebenfalls

$$N = n(\text{RG}) + n(\text{kein RG})$$

gilt. Weiter kann man die Teilmenge der Frauen, die das Risikogewebe haben, aufteilen in die Menge der Frauen, die außerdem an Krebs leiden, und solche, die dies nicht tun:

$$n(\text{RG}) = n(\text{RG und Krebs}) + n(\text{RG und kein Krebs}). \tag{3}$$

Ebenso kann man die Teilmenge der Frauen, die kein Risikogewebe haben, aufteilen in die Menge derjenigen, die an Krebs leiden, und die Menge derjenigen, die nicht an Krebs leiden:

$$n(\text{kein RG}) = n(\text{kein RG und Krebs}) + n(\text{kein RG und kein Krebs}).$$

Die Tabelle 1 verdeutlicht die Beziehungen zwischen diesen Häufigkeiten:

Nun ist es von Interesse, zu wissen, wie groß die Wahrscheinlichkeit ist, Krebs zu bekommen, wenn man das Risikogewebe hat. Die Wahrscheinlichkeit kann man

[5]McGee, G. (1979) Breast surgery before cancer. *The Ann Arbor News*, Feb. 6: B-1. S.a. Dawes, R.M.: Everyday Irrationality. Westview Press 2001

Tabelle 1: Häufigkeiten des Auftretens von Risikogewebe (RG) und Krebs; N Gesamtzahl der Fälle

	Krebs	kein Krebs	Summe
RG	n(RG und Krebs)	n(RG und kein Krebs)	n(RG)
kein RG	n(kein RG und Krebs)	n(kein RG und kein Krebs)	n(kein RG)
Summe	n(Krebs)	n(kein Krebs)	N

durch eine entsprechende relative Häufigkeit abschätzen: so kann man z.B. für eine zufällig gewählte Frau die Wahrscheinlichkeit, dass sie das Risikogewebe hat, durch die relative Häufigkeit

$$h(\text{RG}) = \frac{n(\text{RG})}{N}$$

abschätzen. Die Wahrscheinlichkeit, dass sie das Risikogewebe *und* Krebs hat, schätzt man dementsprechend durch die relative Häufigkeit

$$h(\text{RG und Krebs}) = \frac{n(\text{RG und Krebs})}{N}. \tag{4}$$

h(RG und Krebs) ist also eine Schätzung der Wahrscheinlichkeit des Ereignisses, eine Frau aus der Stichprobe zu wählen, die sowohl das Risikogewebe wie auch Krebs hat. Der Wert von h(RG und Krebs) gibt aber noch nicht die Wahrscheinlichkeit an, an Krebs zu erkranken, wenn man das Risikogewebe RG hat. Dazu muß man die *bedingten Wahrscheinlichkeiten* bzw. deren Abschätzungen, die bedingten relativen Häufigkeiten h(Krebs|RG) und h(Krebs|kein RG) kennen. Diese Wahrscheinlichkeiten bzw. relativen Häufigkeiten geben an, mit welcher Wahrscheinlichkeit eine Frau Krebs bekommt, wenn sie entweder das Risikogewebe RG hat oder nicht hat. Die erstere ist durch den Quotienten

$$h(\text{Krebs}|\text{RG}) = \frac{n(\text{Krebs und RG})}{n(\text{RG})}. \tag{5}$$

definiert, und die zweite dementsprechend durch

$$h(\text{Krebs}|\text{kein RG}) = \frac{n(\text{Krebs und kein RG})}{n(\text{kein RG})}. \tag{6}$$

Man kann diese Quotienten in der folgenden Weise deuten: angenommen, man weiss bereits, dass eine Frau das Risikogewebe hat. Man kann nun fragen, ob sie auch an Krebs leidet. Die Kenntnis über ihr Risikogewebe verändert die Abschätzung der Wahrscheinlichkeit, dass sie auch an Krebs leidet: hat die Frau das Risikogewebe, so ist die Wahrscheinlichkeit höher als wenn sie es nicht hat. Diese Frage ist wichtig für den Fall, dass bei einer Frau (noch) kein Krebs diagnostiziert worden ist. In diesem Fall lautet die Frage, ob sich die Wahrscheinlichkeit, noch an Krebs zu erkranken, erhöht, wenn sie das Risikogewebe hat. Anhand von (5) wird deutlich, dass diese Wahrscheinlichkeit durch den Anteil der Frauen, die das

Risikogewebe und Krebs haben, an der Teilmenge der Frauen mit Risikogewebe ist. Deren Anzahl is $n(\text{RG})$, und diese setzt sich nach (3) aus der Anzahl der Frauen mit Risikogewebe und Krebs plus der Anzahl der Frauen mit Risikogewebe, aber ohne Krebs zusammen.

Es ist wichtig, sich hier den inhaltlichen Unterschied zwischen den hier vorkommenden Größen klar zu machen: sicherlich gilt

$$h(\text{Krebs}|\text{RG}) \neq h(\text{Krebs und RG}),$$

denn

$$\frac{n(\text{Krebs und RG})}{N} \neq \frac{n(\text{Krebs und RG})}{n(\text{RG})}$$

für $N > n(\text{RG})$, wenn also die Anzahl der Frauen mit Risikogwebe kleiner ist als die ohne Risikogewebe. Es folgt insbesondere $h(\text{Krebs}|\text{RG}) > h(\text{Krebs und RG})$.

Man muß nun noch die bedingte relative Häufigkeit $h(\text{RG}|\text{Krebs})$ betrachten, die gemäß

$$h(\text{RG}|\text{Krebs}) = \frac{n(\text{Krebs und RG})}{n(\text{Krebs})} \tag{7}$$

berechnet wird; dies ist der Anteil der Frauen, die sowohl das Risikogewebe wie auch Krebs haben, an der Teilpopulation derjenigen Frauen, die unter Krebs leiden. Vergleicht man nun diese beiden Gleichungen (5) und (7) miteinander, so sieht man, dass sich sich in bezug auf den Nenner unterscheiden: im Ausdruck für $h(\text{Krebs}|\text{RG})$ ist der Nenner $n(\text{RG})$, im Ausdruck für $h(\text{RG}|\text{Krebs})$ ist der Nenner $n(\text{Krebs})$. Sind diese beiden Teilpopulationen ungleich groß, gilt also

$$n(\text{RG}) \neq n(\text{Krebs}),$$

so werden auch die bedingten relativen Häufigkeiten ungleich sein.

Hat eine Frau das Risikogewebe, aber noch keinen Krebs, so kann sie sich fragen, ob sie den Krebs noch bekommen wird. Eine Frau ohne das Risikogewebe und ohne Krebs kann sich das gleiche fragen. Hat sie das Risikogewebe, so gehört sie eben zur Teilpopulation mit dem Risikogewebe, und die Wahrscheinlichkeit für Frauen mit Risikogewebe, aber ohne Krebs, an Krebs zu erkranken, ist durch $h(\text{Krebs}|\text{RG})$ gegeben, d.h. durch den Anteil der Frauen, die Krebs bekommen, wenn sie zur Teilpopulation (-stichprobe) der Frauen mit Risikogewebe gehören. Der Arzt ist aber von einer anderen Teilpopulation ausgegangen: er betrachtete den Anteil der Frauen mit Risikogewebe an der Teilpopulation derjenigen, die an Krebs litten, und zu denen gehören auch diejenigen Frauen, die kein Risikogewebe hatten. Er fand, dass Frauen mit Krebs in größerer Zahl das Risikogewebe hatten, als dies in der Population allgemein vorkommt, d.h. er hat den Anteil $h(\text{Risikogewebe}|\text{Krebs})$ betrachtet. Diesen Anteil hat er dann ungerechtfertigterweise gleich $h(\text{Krebs}|\text{RG})$ gesetzt. Diese Gleichsetzung ist aber nur dann erlaubt, wenn

$$n(\text{RG}) = n(\text{Krebs})$$

gilt, wie der Vergleich der Gleichungen (5) und (7) lehrt! Nun ist es aber so, dass längst nicht alle Frauen mit dem Risikogewebe auch tatsächlich Krebs bekommen,

und dies bedeutet

$$n(\text{RG}) > n(\text{Krebs}),$$

und diese Tatsache wiederum bedeutet, dass

$$h(\text{Krebs}|\text{RG}) < h(\text{RG}|\text{Krebs}), \tag{8}$$

d.h. die Wahrscheinlichkeit, Krebs zu bekommen, wenn man ein Risikogewebe hat, ist kleiner als die, dass man ein Risikogewebe hat, wenn man unter Krebs leidet. Der kalifornische Arzt hat aber Gleichheit dieser bedingten Häufigkeiten angenommen. Dies war sein Fehler; der Fehler der Frauen, die sich einer Mastektomie unterzogen, bestand darin, nicht kritisch nachzurechnen.

Grundquoten: Man muß sich, um die Beziehung zwischen relativen Häufigkeiten und bedingten relativen Häufigkeiten klar zu machen, die Gleichungen (5) und (7) ansehen, die auch erhellen, warum man hier leicht zu Verwechslungen geführt wird. Beide relativen Häufigkeiten sind zum einen durch die Häufigkeit $n(\text{Krebs und RG})$ definiert, d.h. durch die Häufigkeit, mit der Krebs zusammen mit dem Risikogewebe auftritt. Die bedingten relativen Häufigkeiten ergeben sich aber durch Division durch $n(\text{RG})$ bzw. $n(\text{Krebs})$. Der Unterschied zwischen diesen Häufigkeiten bzw. relativen Häufigkeiten ist für die Beurteilung der Abhängigkeit zwischen Merkmalen so wichtig, dass man einen speziellen Ausdruck für sie eingeführt hat:

> Dies relativen Häufigkeiten $n(\text{RG})/N$ und $n(\text{Krebs})/N$, heißen Grundquoten; sie geben an, wie häufig das Risikogewebe bzw. Krebs überhaupt in der Gesamtpopulation auftritt.

Nur wenn diese Grundquoten gleich groß sind, sind auch die bedingten relativen Häufigkeiten gleich groß. Man kann die Gleichungen (5) und (7) zu einer einzigen zusammenfassen:

$$h(\text{Krebs}|\text{RG}) = h(\text{RG}|\text{Krebs})\frac{n(\text{Krebs})}{n(\text{RG})}. \tag{9}$$

Man sieht dann sofort, dass die bedingten relativen Häufigkeiten nur dann gleich sind, wenn $n(\text{Krebs})/n(\text{RG}) = 1$, d.h. wenn beide Häufigkeiten gleich groß sind. Man kann dementsprechend sagen, dass der kalifornische Arzt die Unterschiedlichkeit der Grundquoten vernachlässigt hat. Die Vernachlässigung der Grundquoten impliziert die Vernachlässigung der Asymmetrien der Urteile, d.h. der Schlüsse von Krebs auf Risikogewebe und umgekehrt von Risikogewebe auf Krebs. In der Tabelle 2 sind die relativen Häufigkeiten zusammengefasst worden (die im Beispiel nicht angegbenen relativen Häufigkeiten lassen sich aus den obigen Formeln errechnen). Aus dieser Tabelle läßt sich wiederum errechnen, dass das Risiko einer Frau, an Krebs zu erkranken, wenn sie das Risikogewebe hat, um den Faktor 1.635 höher ist als wenn sie das Risikogewebe nicht hat[6].

[6] Zugrundegelegt wird dabei die Annahme, dass die Wahrscheinlichkeit, an Krebs zu erkranken, durch die logistische Funktion $p(x) = \exp(Ax + B)/(1 + \exp(Ax + b))$ gegeben ist, wobei $x = 1$, wenn das Risikogewebe vorliegt, und $x = 0$, wenn es nicht vorliegt. Das Risiko ist durch $p(x)/(1-p(x))$ definiert, und der Wert 1.365 ergibt sich aus der Maximum-Likelihood-Schätzung für den Parameter A.

Tabelle 2: Risikogewebe und Krebs

	Krebs		
	ja	nein	$\sum$
RG: ja	.0684	.5016	.57
RG: nein	.0331	.3969	.43
$\sum$	.1015	.8985	1.00

Anmerkung: Der vorangegangenen Betrachtung liegt eine Vereinfachung zugrunde: hat eine Frau ein Risikogewebe, so hängt die Abschätzung der Wahrscheinlichkeit, (Brust-)Krebs zu bekommen, auch von dem Zeitpunkt ab, zu dem diese Abschätzung vorgenommen wird. Zu genaueren Abschätzungen gelangt man, wenn man die Wahrscheinlichkeit, zu erkranken, unter der Bedingung, bis jetzt noch nicht erkrankt zu sein, für die beiden Teilpopulationen "Frauen mit RG" und "Frauen ohne RG" betrachtet. Für die Verdeutlichung der Asymmetrie zwischen h(RG|Krebs) und h(Krebs|RG) ist die hier vorgestellte Betrachtung aber zunächst hinreichend. □

Der Schluß von einer bedingten Häufigkeit bzw. Wahrscheinlichkeit auf die dazu korrespondierende inverse (also von $h(A|B)$ auf $h(B|A)$) ist im Alltagsdenken im allgemeinen fehlerhaft, die dabei nötige Berücksichtigung der Grundquoten ist offenbar schwierig und wird dementsprechend von psychologischen, medizinischen und anderen Experten oft mißachtet. So beruht das Argument, Marihuana sei mit großer Warhscheinlichkeit eine Einstiegsdroge, auf dem Sachverhalt, dass Personen, die Heroin spritzen, im allgemeinen auch Cannabis rauchen. Aber p(Cannabis|Heroin) ist eben nicht gleich p(Heroin|Cannabis); diese bedingte Wahrscheinlichkeit ist sehr viel kleiner als p(Cannabis|Heroin). In Bezug auf die Tolerierung des oft sehr viel schädlicheren Alkohols reflektiert die Ächtung des Cannabisrauchens eine typische und oft auftretende Inkonsistenz im Denken[7], die auf der Vernachlässigung von Grundquoten einerseits und Vergleich von Quoten und Wahrscheinlichkeiten generell andererseits zurückgeht und die bei "verstehender" Urteilsbildung i.a. unentdeckt bleibt, aber gleichwohl das Urteil stark verfälschen kann.

Die Konzeption der Psychologie als Naturwissenschaft beruht auf der Grundauffassung, dass Aussagen im Prinzip überprüfbar sein sollen. Dazu gehört, dass Annahmen möglichst explizit gemacht werden und folglich empirische Resultate nur in bezug auf den gewählten Rahmen von (Hintergrunds-)Annahmen interpretiert werden sollen; dass sich auch hier Streit bzw. die von Jüttemann (1991) so gegeißelten "Glaubenskämpfe" bezüglich als selbstverständlich oder eben nicht als selbstverständlich geltende Annahmen ergeben können, ist normal, es gibt sie in jeder Naturwissenschaft. Weiter ergibt sich die Notwendigkeit, die verwendeten Begriffe explizit zu definieren. Dies geschieht oft als sogenannte *Operationalisierung*, d.h. man definiert einen Begriff, z.B. "Stress", durch die Art, wie man die

[7] Ziel dieser Argumentation ist keineswegs, das Rauchen von Cannabis zu fördern!

entsprechende Größe mißt. Dabei ist "Messen" sehr allgemein definiert, auch die bloße Kategorisierung (männlich-weiblich, gestresst-nicht gestresst, etc) gilt als Messung, es müssen allerdings die Kategorien definiert sein (wann gilt eine Person "gestresst", etc?). Die Notwendigkeit der Operationalisierung zieht wiederum eine gewisse Beschränkung der Menge der denkbaren Begriffe nach sich: man wird nicht versuchen, das "Seelische" zu operationalisieren, wenn mit diesem Wort der Begriff gemeint ist, den die "klassische" geisteswissenschaftliche Psychologie ihren Betrachtungen unterlegt hat, denn in diesem Fall ist "das Seelische" per definitionem nicht operationalisierbar.

Natürlich muß man auf den Vorwurf der geisteswissenschaftlichen Vertreter der Psychologie zu sprechen kommen, das Ziel, nomologische Gesetze aufstellen zu wollen setze einen mechanistischen Geist- und Seelenbegriff voraus, denn der Mensch sei ja eben *kein homo nomologicus*. Eine abstrakte Diskussion darüber, ob der Mensch dies nun sei oder nicht, geht ins Uferlose. Im Rahmen dieses Vortrages lassen sich Möglichkeiten und Grenzen des naturwissenschaftlichen Ansatzes am besten an Beispielen illustrieren. Dies gilt auch für den Nutzen der Statistik, wenn es nicht um Grundlagenforschung, sondern um die Evaluation täglicher therapeutischer oder beratender Arbeit geht.

1.3 Gründe für Fehler beim Verstehen und Vorhersagen

Es gibt viele Gründe, deretwegen Diagnosen und vor allem Vorhersagen nicht hundertprozentig korrekt sein können: Menschen können sich wandeln, situative Faktoren können sich verändern, und die Veränderung der sozialen Umgebung kann mit persönlichen Eigenschaften, auch wenn diese konstant bleiben, wechselwirken und auf diese Weise neue Verhaltensweisen erzeugen. Man kann also vermuten, dass es stets eine Obergrenze für die Präzision insbesondere von Vorhersagen von Verhalten gibt. Es bleibt aber zu erklären, warum Menschen, auch wenn sie als Experten mit langer Erfahrung gelten, im allgemeinen nicht besser urteilen als so einfache "mechanische" Formeln wie (**??**). Dawes et al (1969) haben die hierzu existierende Evidenz zusammengefasst; den Stand der Forschungen seit 1969 (es gibt hier keine qualitativ neuen Einsichten, sondern die alten werden immer wieder bestätigt) findet man in Dawes (1996) und Dawes (2001).

1. Situative Faktoren wie Müdigkeit, unmittelbar vorangehende Faktoren ("recency effects"), geringfügige Veränderung in der Konzeptualisierung und Verarbeitung der vorliegenden Information etc erzeugen zufällige Schwankungen im Urteilsprozess. Dies äußert sich in einer reduzierten "Zuverlässigkeit" (*Reliabilität*) und damit Genauigkeit der Urteile,

2. Ein Symptom heißt *valide*, wenn es das vorherzusagende Merkmal tatsächlich anzeigt. Die Untersuchungen zum Urteilsverhalten von Menschen legen nahe, dass auch erfahrene Personen immer Schwierigkeiten haben, *valide* und *nicht valide* Symptome oder Merkmale zu unterscheiden. Klinische Psychologen und Psychiater erhalten in der Praxis oft keine genaue Rückkopplung über die Exaktheit ihrer Diagnosen, so dass "Erfahrung" konstituierende Lernvorgänge gar nicht stattfinden, weil es eben keine Rückkopplung gibt. Dawes

(1996) zitiert eine Psychotherapeutin, die angibt, ihre Erfahrung mit durch sexuellen Mißbrauch psychisch belasteten Frauen erlaube es ihr, sexuell mißbrauchte Frauen *an ihrem Gang erkennen zu können*. Die Therapeutin ist Opfer dieser mangelnden Rückkopplung.

3. Klinische Urteile können sich selbst erfüllende Prophezeiungen erzeugen. Ein Beispiel mag dies erläutern: in einem Mordprozess wurde ein Angeklagter von einem Psychiater als auch in Zukunft gewalttätig beurteilt. Der Angeklagte wurde daraufhin zum Tode verurteilt. In der Todeszelle verhielt sich der Verurteilte gewalttätig und schien deshalb die Beurteilung durch den Psychiater zu bestätigen. Andererseits hatte er nichts mehr zu verlieren; bei einer anderen Beurteilung und einer dementsprechend anderen Verurteilung hätte sein Verhalten völlig anders sein können.

4. Ein Verhalten erscheint im allgemeinen, wenn es erst einmal eingetreten ist, als vorhersagbarer als zu dem Zeitpunkt, zu dem es noch vorhergesagt werden muß. Bereits gestellte Diagnosen erscheinen daher subjektiv konsistent mit den tatsächlichen Befunden zu sein. Die tatsächlichen Befunde haben auf diese Weise keine korrigierende Wirkung; die *Erfahrung* wird also nicht vermehrt. *Vor* der Diagnose durch einen Experten kann allerdings erhebliche Unsicherheit herrschen (Arkes et al., 1981).

5. Insbesondere Kliniker werden gewissermaßen Opfer ihres Berufes, weil sie beruflich mit einer speziellen Auswahl aus der Bevölkerung, nicht mit einer insgesamt repräsentativen Stichprobe aus ihr zu tun haben. Die Beziehungen zwischen Symptomen und Merkmalen und den zu diagnostizierenden bzw. vorauszusagenden Merkmalen stellen sich deshalb verzerrt dar. Spreen (1981) (zitiert nach Dawes et al., 1989) berichtet z.B. dass die Hälfte der Jugendlichen, die wegen irgendwelcher Straftaten aufgefallen sind, im EEG leichte Abweichungen von als "normal" geltenden EEG-Strukturen zeigen. Also werden diese Abweichungen als Indikator für jugendliche Delinquenz gewertet. Tatsächlich kommen diese Abweichungen aber eben auch bei der Hälfte der nicht delinquent gewordenen Jugendlichen vor, - nur werden diese eben gar nicht erst untersucht; tatsächlich sind derartige Abweichungen bei "normalen", also nicht delinquenten Kindern und Jugendlichen ganz "normal" und geben also keinerlei Hinweis auf zu erwartende Delinquenz. Gleichwohl: hat man erst einmal eine Hypothese gebildet, so neigt man dazu, sie gegen widersprechende Fakten zu *immunisieren*, sodass die Konsistenz der "Erfahrung" mit der Hypothese überschätzt wird. Dementsprechend wird die Gültigkeit der widersprechenden Information *unterschätzt*. Dieses Phänomen ist allgemein als urteilsverzerrender Faktor unter dem Namen

6. *Repräsentativität* (representiveness) bekannt. Der beurteilende Experte oder Diagnostiker bezieht sich bei seiner Beurteilung auf die Übereinstimmung einiger beobachteter Merkmale (z.B. EEG-Abweichungen) mit stereotypen Kategorien in seinem Gedächtnis, ohne die jeweiligen bedingten Wahrscheinlichkeiten zu berücksichtigen. Tversky und Kahneman (1974) betrachten das folgende Beispiel: Eine Person wird als scheu und zurückgezogen mit einer

Neigung zu Ordnung und Detail beschrieben. Wie groß ist die Wahrscheinlichkeit, dass sie diese Person als (a) Bauer, (b) Handelsvertreter, (c) Pilot eines Verkehrsflugzeuges, (d) Bibliothekar, oder (e) Arzt einschätzen? Das Urteil wird dann, der Repräsentativitätsstrategie entsprechend, nach Maßgabe der Ähnlichkeit der Beschreibung der Person mit den Stereotypen der Berufe getroffen (z.B. "wahrscheinlich ist der Mann Bibliothekar"), nicht aber nach Maßgabe der relativen bedingten Häufigkeit (bedingte Wahrscheinlichkeit), mit der Angehörige der verschiedenen Berufe die beobachteten Eigenschaften tatsächlich haben.

Ein anderer verzerrender Einfluß auf die Urteile nicht nur von "normalen" Menschen, sondern auch von Experten ist die

7. *Verfügbarkeit* (Availability) (Tversky et al., 1974). Hier wird die Wahrscheinlichkeit, mit der eine Person ein Merkmal hat, in Abhängigkeit von der Anzahl von Beispielen, die man dafür im Gedächtnis hat beurteilt. Das Risiko von Herzinfarkten von Menschen im mittleren Alter wird nach Maßgabe der Häufigkeit, mit der in der eigenen Bekanntschaft solche Herzinfarkte aufgetreten sind, beurteilt. Die Schätzung hängt nun aber davon ab, wie gut man solche Beispiele erinnern kann: Herzinfarkte sind wegen ihrer drastischen Konsequenzen leichter zu erinnern, als weniger saliente Merkmale. In jedem Fall kommt es hier zu Fehlabschätzungen, die zu drastischen Fehlbeurteilungen führen können.

Literatur

[1] Anastasi, A.; Differential psychology: individual and group differences in behavior. New Yord 1964

[2] Arkes et al., H.R. (1981) Hindsight bias among physicians weighing the likelihood of diagnoses. *Journal of Applied Psychology*, 66, 252–254

[3] Billingsley, P.: Probability and measure. John Wiley & Sons, New York 1979

[4] Bloom, R.F., Brundage, E.G. (1947) Predictions of success in Elementary School for Enlisted Personal. IN: D.B. Stuit, ed. Personal Research and Test Development in the National Bureau of Personnel, 233-261, Princeton,N.J.

[5] Breuer, F.: Einführung in die Wisenschaftstheorie für Psychologen, Münster, 1991.

[6] Chapman, L.J., Chapman, J.P. (1967) Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, 193 - 204

[7] Chapman, L.J., Chapman, J.P. (1969) Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271–280

[8] Dawes, R.M.: House of Cards. Psychology and psychotherapy built on myth. New York 1996

[9] Dawes, R.M.: Everyday Irrationality. Westview Press 2001

[10] Dawes, R.M., Faust, D., Meehl, P.E. (1989) Clinical versus actuarial judgment, Science, 243, 1668-1674

[11] Dilthey, W.: Ideen über eine beschreibende und zergliedernde Psychologie (1904), in : Dilthey, W.: Die Philosophie des Lebens. Aus seinen Schriften ausgewählt von Herman Nohl, Göttingen 1961

[12] DeVaul, R.A., Jervey, F., Chappell, J.A., Carver, P., Short, B., O'Keefe, S. Medical School Performance of initially rejected students. Journal of the American MEdical Association, 257, 1986, 47-51

[13] Dollard, J., Doob, L.W., Miller, N.E., Mowrer, O.H., Sears, R.R., Ford, C.S., Hovland, C.I., Sollenberger, R.T.: Frustration and Aggression. New Haven: Yale University Press 1939

[14] Driesch, H.: Grundprobleme der Psychologie. Ihre Krisis in der Gegenwart. Leipzig 1929

[15] Einhorn, H.J.: (1972) Expert measurement and mechanical combination. Organizational Behaviour and Human Performance, 7, 86-106

[16] Einhorn, J., Hogarth, R.M. (1978) Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 85, 395 – 416

[17] Gadamer, H.-G. Wahrheit und Methode. Tübingen, 1965

[18] Habermas, J.: Der deutsche Idealismus der jüdischen Philosophen. In: Koch. T.: Porträts zur deutsch-jüdischen Geistesgeschichte. Köln, 1997

[19] Feynman R. Cargo cult science. In: Surely you're joking Mr. Feynman! London, Unwin Paperbacks, 1985.

[20] Geuter, U. Nationalsozialistische Ideologie und Psychologie. In: Ash, M.G., Geuter, U.: Geschichte der deutschen Psychologie im 20. Jahrhundert. Opladen, 1985

[21] Goldberg, L.R. (1960) Simple models or simple processes? Some research on clinical judgments. American Psychologist, 23, 483-496

[22] Hofstätter, P.R.: Differentielle Psychologie. Stuttgart 1971

[23] Humboldt, W. v.: Über die männliche und weibliche Form, Berlin 1795

[24] Husserl, E. Logische Untersuchungen. Halle 1913, Cartesianische Meditationen und Pariser Vorträge, Haag 1950

[25] Jaeggi, E.: Die Forschungsvignette und ihr Bezug zum psychoanalytischen Denken. In: Schorr, A.: Die Psychologie und die Methodenfrage. Hogrefe-Verlag, Göttingen 1994

[26] Jüttemann, G. (1991) Zwischen Wissenschaft und Glaubenslehre: Psychologie ohne Identität. Report Psychologie, April-Ausgabe, 19-24

[27] Klages, L.: Der Geist als Widersacher der Seele. München und Bonn, 1960

[28] Klages, L.: Einführung zu Schuler: Fragmente und Vorträge aus dem Nachlaß. Leipzig, 1940

[29] Klemm. O.: Geschichte der Psychologie. Leipzig, 1911

[30] Lakatos, I: Falsification and the methodology of scientific research programs. In: Lakatos, I.. Musgrave, A. (edts) Criticsm and the Growth of Knowledge. Cambridge University Press, Cambridge, 1970.

[31] Legewie, H. (1991): Argumente für eine Erneuerung der Psychologie. Report Psychologie, Februar-Ausgabe, 11-20

[32] Lersch, P.; Aufbau der Person. München 1966

[33] Lersch, P.: Seele und Welt. Zur Frage nach der Eigenart des Seelischen. Leipzig 1941

[34] Meehl, P. (1954) Clinical versus statistical prediction: A theoretical analysis and review of the literature, University of Minnesota Press

[35] Milstein, R.M., Wilkinson, L., Burrow, G.N., Kessen, W. (1981) Admission decisions and performance during Medical School. Journal of Medical Education, 56, 77-82

[36] Monahan, J. (1984) The prediction of violent behavior: toward a second generation of theory and policy. American Journal of Psychiatry, 141, 10-15

[37] Penrose, R.: The emperors new mind. Oxford University Press, Oxford 1989

[38] Sawyer, J. (1966) Measurement and prediction, Clinical and Statistical. Psychological Bulletin, 66, 178-200

[39] Scabó, I.: Geschichte der mechanischen Prinzipien. Birkhäuser Verlag, Basel und Stuttgart, 1976

[40] Schiwkoff, G.: Philosophisches Wörterbuch. Kröner Verlag, Stuttgart 1957

[41] Seifert, F.: Charakterologie. Handbuch der Philosophie, München-Berlin 1929

[42] Smith, M.L., Glass, G.V. (1977) Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760

[43] Spengler, O. Der Untergang des Abendlandes. Umrisse einer Morphologie der Weltgeschichte. Lizenzausgabe in einem Band für Ex Libris, Zürich 1980.

[44] Spranger, E.: Lebensformen. Geisteswissenschaftliche Psychologie und Ethik der Persönlichkeit. Tübingen, 1959

[45] Stegmüller, W.: Hauptströmungen der Gegenwartsphilosophie, Band II, Stuttgart 1987

[46] Stegmüller, W.: Erklärung, Begründung, Kausalität. Berlin, 1983

[47] Stegmüller, W.: Das Problem der Induktion: Humes Herausforderung und moderne Antworten. Der sogenannte Zirkel des Verstehens. Wissenschaftliche Buchgesellschaft Darmstadt, Darmstad, 1974

[48] Stein, D.M., Lambert, M.J. (1984) On the relationship between therapist experience and psychotherapy outcome. Clinical Psychology Review, 4, 127-142

[49] Szabo, I. Geschichte der mechanischen Prinzipien. Stuttgart 1976

[50] Tversky, A., Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

[51] Webster, R.: Why Freud was wrong. London 1995

[52] Weitbrecht, H.J.: Psychiatrie im Grundriss. Berlin 1963

[53] Wellek, A.: Der Rückfall in die Methodenkrise der Psychologie und ihre Überwindung. Göttingen 1959

[54] Wellek, A.: Die Polarität im Aufbau des Charakters. System der konkreten Charakterkunde. Bern 1966

[55] Westphal, K. (1931) Körperbau und Charakter der Epileptiker. Nervenarzt, 4, 96-99

[56] Ziege, E.M.: Die "Mörder der Göttinnen". In: Schoeps, J.H., Schlör, J.: Antisemitismus. Vorurteile und Mythen. Piper, München, Zürich, 1995