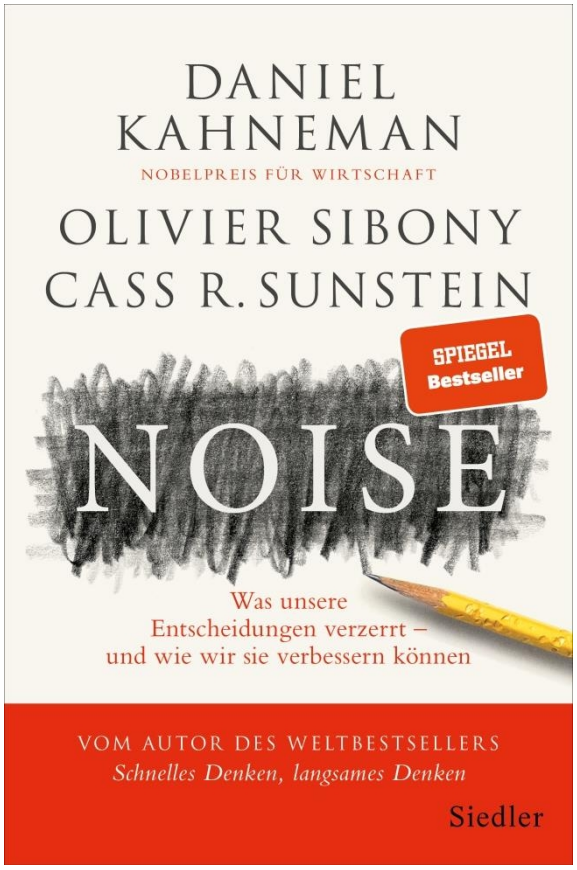



# Leseprobe

Daniel Kahneman, Olivier Sibony, Cass R. Sunstein

**Noise**

Was unsere Entscheidungen verzerrt – und wie wir sie verbessern können

»Voller erhellender Beispiele. Das Buch hat das Potential, zu einem Standardwerk für Entscheider zu werden, für Personalabteilungen genauso wie für Ingenieure, Sozialarbeiter oder Fußballtrainer.« *Frankfurter Allgemeine Zeitung*



Seiten: 480

Erscheinungstermin: 17. Mai 2021

Mehr Informationen zum Buch gibt es auf

www.penguinrandomhouse.de

# Inhalte



# Zum Buch

**Das neue Buch des Nobelpreisträgers Daniel Kahneman, Autor des Weltbestsellers »Schnelles Denken, langsames Denken«: nominiert für den Deutschen Wirtschaftsbuchpreis 2021**

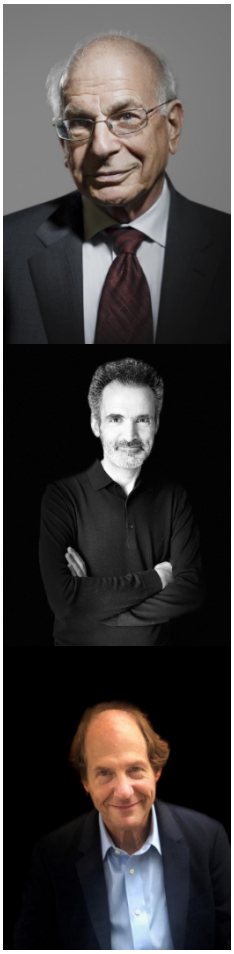
Warum treffen wir, je nach Umständen, völlig unterschiedliche Entscheidungen auf ein und derselben Faktengrundlage?

Wieso kommen zwei Experten, die über identische Informationen verfügen, zu komplett anderen Schlussfolgerungen?

Weshalb entscheiden wir uns immer wieder falsch, ob im Beruf oder im Privatleben?

In seinem neuen Buch, das in Zusammenarbeit mit Bestsellerautor Cass Sunstein und Olivier Sibony entstanden ist, klärt Nobelpreisträger Daniel Kahneman über die Vielzahl von oft zufälligen Faktoren auf, die unsere Entscheidungsfindung stören und häufig negativ beeinflussen – sie sind im Begriff »Noise« zusammengefasst. Wir müssen lernen, diese »Störgeräusche« zu verstehen und mit ihnen umzugehen, nur dann können wir auf Dauer bessere Entscheidungen treffen.

**Dieses Buch ist ein Meilenstein zum Verständnis der Grundlagen unseres Handelns und gehört schon jetzt mit seinem zeitlosen Klassiker »Schnelles Denken, langsames Denken« zur Pflichtlektüre für Entscheidungsträger.**



**Autor**

# Daniel Kahneman, Olivier Sibony, Cass R. Sunstein

Daniel Kahneman, geboren 1934 in Tel Aviv, ist einer der weltweit einflussreichsten Kognitionspsychologen. Nach Stationen an der Hebrew University in Jerusalem und der University of British Columbia war er bis 1994 Professor an der University of California in Berkeley und hat seither die Eugene-Higgins-Professur für Psychologie an der Woodrow Wilson School der Princeton University inne. Kahneman revolutionierte die Wissenschaft vom menschlichen Verhalten, indem er die Erkenntnisse der Hirnforschung und der Verhaltensbiologie zusammenführt und auf die Wirtschaftswissenschaften anwendet. Für seine Arbeit erhielt Kahneman zahlreiche Auszeichnungen namhafter Universitäten und wurde 2002 mit dem Wirtschaftsnobelpreis ausgezeichnet. »Schnelles Denken, langsames Denken« wurde zum Weltbestseller und rangiert seit vielen Jahren ganz oben in den Bestsellerlisten.

Cass R. Sunstein, geboren 1954, ist Jurist und Inhaber des Felix-Frankfurter-Lehrstuhls an der Harvard Law School. Er war Berater von Barack Obama zu Intelligence and Communications Technologies und ist Autor zahlreicher Bücher, darunter »The World According to Star Wars« und »Nudge. Wie man kluge Entscheidungen anstößt« (mit Richard Thaler), das zum Bestseller wurde.

Olivier Sibony ist Autor, Dozent und Unternehmensberater, spezialisiert auf strategische Entscheidungsfindung und die Organisation von Entscheidungsprozessen. Er arbeitete 25 Jahre als Consultant, Partner und Direktor bei McKinsey & Company in Paris, New York und Brüssel. Als Affiliate Professor an der Business School HEC in Paris

Daniel Kahneman, Olivier Sibony, Cass R. Sunstein

# NOISE

*Für Noga, Ori und Gili – DK*
*Für Fantin und Lélia – OS*
*Für Samantha – CRS*

# Inhalt

# EINLEITUNG
# Zwei Arten von Fehlern

Stellen Sie sich vor, vier Gruppen von Freunden gehen zu einem Schießstand. Jede Gruppe besteht aus fünf Personen; sie benutzen gemeinsam ein Gewehr, und jede Person schießt einmal.

Abbildung 1 zeigt die Ergebnisse.

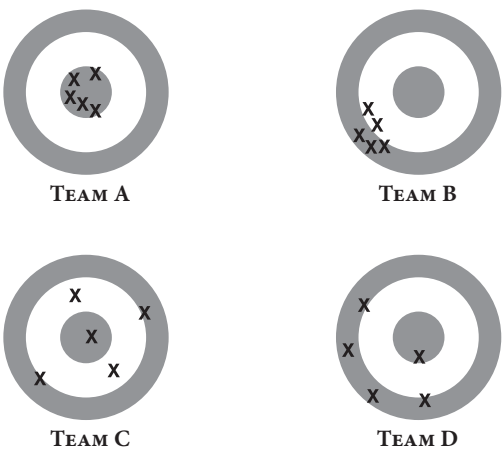



**Abbildung 1: Vier Teams**

In einer idealen Welt wäre jeder Schuss ein Volltreffer.

Das ist bei Team A fast der Fall. Die Treffer ballen sich dicht im Schwarzen zusammen, in der Mitte der Zielscheibe, und bilden ein fast perfektes Muster.

Das Resultat von Team B nennen wir *biased* (»verzerrt«), weil es systema-

tisch danebengeschossen hat. Wie aus der Abbildung zu ersehen ist, lässt sich aus der Konsistenz des Bias, in diesem Fall also der Zielabweichung, eine Vorhersage ableiten. Würde eines der Mitglieder des Teams ein weiteres Mal schießen, würden wir darauf wetten, dass der Treffer im gleichen engen, abweichenden Bereich wie die ersten fünf läge. Die Beständigkeit des Bias legt auch eine kausale Erklärung nahe: Vielleicht war das Zielfernrohr am Gewehr des Teams verbogen.

Das Ergebnis von Team C nennen wir *noisy* (»verrauscht«), weil die Treffer breit gestreut sind. Es gibt kein offensichtliches Bias, weil die meisten Einschüsse in grober Näherung auf einem Kreis um die Mitte der Scheibe liegen. Wenn eines der Mitglieder des Teams einen weiteren Schuss abgeben würde, könnten wir kaum abschätzen, wo genau der Treffer landen würde. Außerdem fällt einem zur Erklärung der Ergebnisse von Team C keine interessante Hypothese ein. Wir wissen, dass vier seiner Mitglieder schlechte Schützen sind. Wir wissen nicht, warum ihre Treffer so verrauscht, so breit gestreut sind.

Das Resultat von Team D ist sowohl verzerrt als auch verrauscht. Vergleichbar mit Team B haben seine Mitglieder systematisch nicht die Mitte der Zielscheibe getroffen, und wie bei Team C sind die Treffer breit gestreut.

Aber dies ist kein Buch über das Schießen auf Zielscheiben. Unser Thema sind Urteilsfehler. Bias und Noise – systematische Abweichung und Zufallsstreuung – sind verschiedene Komponenten von Urteilsfehlern. Die Zielscheiben verdeutlichen den Unterschied.[1]

Der Schießstand ist eine Metapher für das, was bei der Urteilsbildung und insbesondere bei den vielfältigen Entscheidungen, die Menschen in Organisationen, Institutionen oder Unternehmen treffen, schiefgehen kann. In diesen Situationen finden wir die beiden Arten von Fehlern, die in Abbildung 1 veranschaulicht werden. Manche Urteile sind verzerrt; sie liegen systematisch »daneben«. Andere Urteile sind verrauscht, das heißt, sie sind weit um »das Ziel« gestreut, obwohl sie eigentlich übereinstimmen sollten. Leider sind viele Organisationen sowohl von Bias als auch von Noise betroffen.

Abbildung 2 veranschaulicht einen wichtigen Unterschied zwischen Bias und Noise. Sie zeigt das, was Sie auf dem Schießstand sehen würden, wenn Ihnen nur die *Rückseiten* der Zielscheiben gezeigt würden, auf welche die Teams geschossen haben, ohne dass Sie den geringsten Hinweis darauf hätten, wo sich die Zielscheibenmitte befindet, die von den Schützen anvisiert wurde.

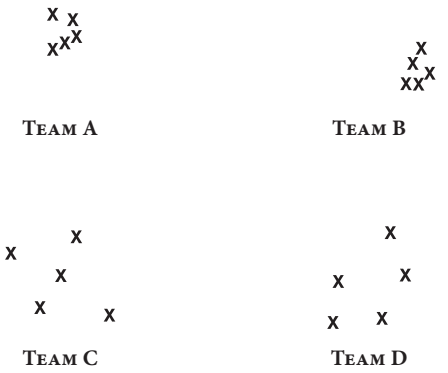



**Abbildung 2: Ein Blick auf die Rückseite der Zielscheiben**

Betrachtet man nur die Rückseite der Zielscheiben, lässt sich nicht sagen, ob Team A oder Team B treffsicherer war. Aber man kann auf den ersten Blick sagen, dass die Treffer der Teams C und D breit gestreut – verrauscht – sind, während dies bei den Teams A und B nicht der Fall ist. Tatsächlich wissen wir über die Streuung genauso viel wie in Abbildung 1. Es ist eine allgemeine Eigenschaft von Noise, dass man es erkennen und messen kann, auch wenn man nichts über das Ziel oder das Bias weiß.

Diese allgemeine Eigenschaft von Noise ist für das, worum es uns in diesem Buch geht, von zentraler Bedeutung, weil viele unserer Schlussfolgerungen auf Urteilen beruhen, bei denen die Wahrheit unbekannt oder sogar unerkennbar ist. Wenn Ärztinnen und Ärzte bei demselben Patienten verschiedene Diagnosen stellen, können wir ihren Dissens erforschen, ohne zu wissen, woran der Patient wirklich leidet. Wenn die Manager einer Filmproduktionsfirma das Marktpotenzial für einen Film abschätzen, können wir die Streuung ihrer Antworten analysieren, ohne zu wissen, wie viel der Film letztendlich eingespielt hat oder ob er überhaupt produziert wurde. Wir müssen nicht wissen, wer recht hat, um zu messen, wie sehr Urteile über denselben Sachverhalt voneinander abweichen. Um Noise zu messen, müssen wir lediglich die Rückseite der Zielscheibe betrachten.

Wenn wir Urteilsfehler verstehen wollen, müssen wir sowohl Bias – die systematische Abweichung, die Verzerrung – als auch Noise – die Zufallsstreuung, das störende Rauschen – verstehen. Wie wir sehen werden, ist Noise

manchmal das wichtigere Problem. Aber in öffentlichen Diskussionen über Urteilsfehler und in Organisationen überall auf der Welt wird dies nur selten erkannt. Bias ist sozusagen der Star der Show, während Noise im Allgemeinen hinter den Kulissen verborgen bleibt. Das Thema systematische Abweichung wurde in Tausenden wissenschaftlichen Aufsätzen und Dutzenden populärwissenschaftlichen Büchern erörtert, von denen nur wenige das Problem der Zufallsstreuung überhaupt erwähnen. Mit diesem Buch versuchen wir, die Dinge wieder ins rechte Lot zu bringen.

Entscheidungsfindungen sind in vielen Lebensbereichen oft durch ein geradezu skandalös hohes Maß an Noise gekennzeichnet. Nachfolgend ein paar Beispiele für das Ausmaß an Zufallsstreuung in Situationen, in denen es auf Treffgenauigkeit ankommt:

- *Medizin:* In Bezug auf denselben Patienten stellen Ärztinnen und Ärzte oft keine einheitlichen Diagnosen; dies betrifft Hautkrebs, Brustkrebs, Herzkrankheiten, Tuberkulose, Lungenentzündungen, Depressionen und viele weitere Erkrankungen. Besonders stark ist Noise in der Psychiatrie, wo subjektive Einschätzungen bei der Diagnose offensichtlich eine wichtige Rolle spielen. Aber auch in Bereichen, in denen man es nicht erwarten würde, wie etwa bei der Interpretation von Röntgenaufnahmen, findet man ein erhebliches Maß an Zufallsstreuung.
- *Entscheidungen über die behördliche Inobhutnahme von Kindern:*[2] Fallmanager in Jugendämtern müssen beurteilen, ob bei einem Minderjährigen eine Kindeswohlgefährdung vorliegt und – falls dem so ist – entscheiden, ob eine Unterbringung in einer Pflegefamilie angezeigt ist. Das System ist durch Noise gestört, da manche Fallmanager viel eher dazu neigen, Kinder in einer Pflegefamilie unterzubringen als ihre Kollegen. Von diesen unglücklichen Minderjährigen, die durch die strengen Sozialarbeiter an Pflegeeltern übergeben wurden, hat dann in späteren Jahren eine Mehrzahl im Leben deutlich weniger erreicht als Nichtpflegekinder, sie werden viel häufiger straffällig, die Mädchen unter ihnen werden häufiger im Teenageralter schwanger, und sie verdienen weniger.
- *Vorhersagen:* Professionelle Prognostiker treffen höchst unterschiedliche Vorhersagen über den wahrscheinlichen Absatz eines neuen Produkts, die wahrscheinliche Zunahme der Arbeitslosenquote, die Wahrschein-

lichkeit, dass angeschlagene Unternehmen pleitegehen, und vieles andere mehr. Aber sie sind nicht nur untereinander uneins, sie stimmen auch mit sich selbst nicht überein. Als zum Beispiel dieselben Softwareentwickler an zwei verschiedenen Tagen gebeten wurden, die Zeit abzuschätzen, die sie bräuchten, um dieselbe Aufgabe zu erledigen, unterschieden sich die von ihnen angesetzten Stundenzahlen durchschnittlich um 71 Prozent.[3]

– *Asylentscheidungen:* Ob ein Asylbewerber in den Vereinigten Staaten anerkannt wird, kommt einer Art Glücksspiel gleich. Bei einer Studie über Fälle, die zufallsabhängig verschiedenen Richtern zugewiesen wurden, kam heraus, dass ein Richter 5 Prozent der Asylsuchenden anerkannte, während ein anderer 88 Prozent anerkannte. Der Titel der Studie sagt alles: »Flüchtlingsroulette«.[4] (Wir werden nachfolgend einiges an Roulette erleben.)
– *Personalentscheidungen:* Personalverantwortliche, die Vorstellungsgespräche führen, schätzen dieselben Bewerberinnen und Bewerber sehr unterschiedlich ein. Auch werden die Leistungen derselben Mitarbeiter höchst unterschiedlich bewertet, und die Bewertung hängt stärker von der Person des Beurteilenden als von der zu beurteilenden Leistung ab.
– *Kautionsentscheidungen:* Ob ein Beschuldigter in den USA gegen Kautionszahlung auf freiem Fuß bleibt oder aber bis zum Prozess in Haft genommen wird, hängt weitgehend von der Person des Richters/der Richterin ab, der/die über die Sache verhandelt. Einige Richter/Richterinnen sind viel nachsichtiger als andere. Auch in Bezug auf die Frage, bei welchen Angeklagten die höchste Flucht- beziehungsweise Rückfallgefahr besteht, kommen Richter und Richterinnen zu sehr unterschiedlichen Einschätzungen.
– *Forensik (Kriminaltechnik):* Uns wurde beigebracht, die Identifikation per Fingerabdruck für absolut sicher zu halten. Aber Sachverständige für Daktyloskopie, die beurteilen sollen, ob ein an einem Tatort gefundener Fingerabdruck eindeutig einem Verdächtigen zugeordnet werden kann, kommen gelegentlich zu unterschiedlichen Schlussfolgerungen. Aber nicht genug damit, dass sich die Experten uneins sind, hinzu kommt, dass dieselben Sachverständigen, wenn ihnen zu verschiedenen Zeitpunkten derselbe Fingerabdruck vorgelegt wird, mitunter widersprüchliche Einschätzungen abgeben. Eine ähnliche Uneinheitlichkeit der Beurteilungen ist auch für andere forensische Disziplinen, sogar für die DNA-Analyse, nachgewiesen.

– *Patentgewährung:* Die Autoren einer führenden Studie über Patentanmeldungen weisen ebenfalls auf das damit verbundene Noise hin: »Ob das Patentamt ein Patent gewährt oder ablehnt, hängt in erheblichem Maße davon ab, welcher Prüfer den Antrag zufälligerweise auf seinen Tisch bekommt.«[5] Unter Gerechtigkeitsaspekten ist diese Uneinheitlichkeit der Urteilsbildung ziemlich beunruhigend.

All diese von »Störgeräuschen« geprägten Situationen sind nur die Spitze eines riesigen Eisbergs. Ganz gleich, welche Beurteilungen man sich näher ansieht, man findet höchstwahrscheinlich Noise. Um die Qualität unserer Urteile zu verbessern, müssen wir sowohl Noise als auch Bias reduzieren.

Dieses Buch hat sechs Teile. In Teil 1 befassen wir uns mit dem Unterschied zwischen Noise und Bias, wir zeigen, dass die Entscheidungsfindung sowohl öffentlicher als auch privater Organisationen verrauscht sein kann, manchmal in einem schockierenden Ausmaß. Um die Dimension des Problems zu verdeutlichen, beginnen wir mit Urteilen auf zwei Gebieten. Beim ersten geht es um strafrechtliche Verurteilungen (also den staatlichen Bereich), beim zweiten um Versicherungen (also den privaten Sektor). Auf den ersten Blick könnten die beiden unterschiedlicher nicht sein. Aber in Bezug auf Noise haben sie viel gemeinsam. Um dies nachzuweisen, führen wir das Konzept des »Noise Audits« ein. Ein Noise Audit soll messen, wie groß das Ausmaß der Nichtübereinstimmung unter Fachkräften ist, die innerhalb einer Organisation die gleichen Fälle bearbeiten.

In Teil 2 befassen wir uns eingehend mit den wesentlichen Merkmalen der Urteilsbildung und mit der Frage, wie man Genauigkeit beziehungsweise Ungenauigkeit (Fehler) messen kann. Urteile unterliegen sowohl Verzerrung als auch Rauschen. Wir beschreiben eine bemerkenswerte Äquivalenz der Auswirkungen der beiden Fehlertypen. Das, was wir »Occasion-Noise« (situatives Rauschen) nennen, ist die Streuung der Urteile über denselben Fall durch dieselbe Person oder Gruppe bei verschiedenen Gelegenheiten. In Gruppendiskussionen entsteht aufgrund vermeintlich belangloser Faktoren wie der Reihenfolge der Sprecher ein erhebliches Maß an Noise.

Teil 3 betrachtet einen Typ von Urteilen, der schon gründlich erforscht wurde, etwas genauer: prädiktive Urteile. Wir gehen auf den wichtigsten Vorteil ein, den Regeln, Formeln und Algorithmen gegenüber der menschlichen

Intuition haben, wenn es um *Vorhersagen* geht: Entgegen der landläufigen Meinung besteht dieser nicht so sehr darin, dass Regeln verlässlichere Erkenntnisse liefern, als vielmehr darin, dass sie nicht durch Noise gestört sind. Wir sprechen über eine grundsätzliche Grenze bei der Qualität prädiktiver Urteile – die objektive Unwissenheit über zukünftige Ereignisse – und darüber, wie diese in Verbindung mit Noise die Qualität der Vorhersage einschränkt. Schließlich wenden wir uns einer Frage zu, die Sie sich höchstwahrscheinlich schon selbst gestellt haben: Wenn Noise mit seinem Rauschen so allgegenwärtig ist, wieso ist es mir dann nicht schon früher aufgefallen?

Teil 4 befasst sich näher mit der Psychologie von Noise. Wir erläutern seine wichtigsten Ursachen. Dazu gehören Unterschiede zwischen Menschen, die auf eine Vielzahl von Faktoren zurückzuführen sind, unter anderem Persönlichkeit und Denkstil, höchst individuelle Variationen in der Gewichtung verschiedener Gesichtspunkte und die Tatsache, dass Menschen die gleichen Skalen in unterschiedlicher Weise anwenden. Wir gehen der Frage nach, warum wir Rauschen nicht wahrnehmen und uns Ereignisse und Urteile, die wir kaum vorhersehen konnten, häufig dennoch nicht überraschen.

Teil 5 beschäftigt sich mit der praktischen Frage, wie wir unsere Urteilsbildung verbessern und Fehler vermeiden können. (Leserinnen und Leser, die sich hauptsächlich für die praktischen Anwendungen der Verringerung von Noise interessieren, können die Diskussion über die Schwierigkeiten von Vorhersagen und die Psychologie der Urteilsbildung in den Teilen 3 und 4 überspringen und direkt mit diesem Teil weitermachen.) Wir sehen uns an, was in der Medizin, in der Wirtschaft, im Bildungswesen, bei Behörden und an anderer Stelle unternommen wird, um Störgeräusche bei der Entscheidungsfindung zu unterdrücken. Wir stellen unter dem Oberbegriff der »Entscheidungshygiene« eine Reihe von Verfahren zur Noise-Reduktion vor. Auch präsentieren wir fünf Fallstudien über Bereiche, in denen Noise erwiesenermaßen eine große Rolle spielt und schon seit Längerem erhebliche Anstrengungen unternommen werden, die Störgeräusche zu reduzieren – interessanterweise mit unterschiedlichem Erfolg. In diesen Fallstudien geht es um unzuverlässige medizinische Diagnosen, Leistungsbeurteilungen, Forensik (Kriminaltechnik), Personaleinstellungen und das Erstellen von Prognosen im Allgemeinen. Zum Abschluss stellen wir ein Verfahren vor, das wir »Strukturiertes Entscheidungsprotokoll« nennen: ein universell einsetzbares Verfahren zur Bewertung von Handlungs-

optionen, das mehrere Schlüsselmethoden der Entscheidungshygiene umfasst und seine Anwender in die Lage versetzen soll, weniger mit Noise behaftete, zuverlässigere Urteile zu treffen.

Was ist das wünschenswerte Ausmaß an Noise? Teil 6 wendet sich dieser Frage zu. Auch wenn es der Intuition widersprechen mag, ist das wünschenswerte Ausmaß an Rauschen nicht gleich null. In einigen Bereichen ist es schlichtweg nicht machbar, Noise vollständig zu unterdrücken. In anderen Bereichen wäre es zu kostspielig. Auf wieder anderen Gebieten würden Bemühungen zur Beseitigung von Störgeräuschen andere wichtige Werte gefährden; so könnten solche Maßnahmen zum Beispiel die Arbeitsmoral untergraben und Menschen das Gefühl geben, dass sie als bloße Rädchen in einem Getriebe behandelt werden. Algorithmen können hier hilfreich sein, wecken aber eine Reihe von Bedenken; auf einige davon gehen wir hier ein. Dennoch ist das gegenwärtige Ausmaß an Noise nicht akzeptabel. Wir empfehlen privaten und öffentlichen Organisationen dringend, Noise Audits durchzuführen und sich intensiver als bislang darum zu bemühen, diese Störgeräusche zu beseitigen. Auf diese Weise könnten Organisationen dazu beitragen, weitverbreitete Ungerechtigkeiten abzumildern – und in vielen Bereichen Kosten zu senken.

Dieses Ziel vor Augen, beschließen wir jedes Kapitel mit einigen kurzen Aussagen in Form von Zitaten. Sie können diese Aussagen, so wie sie sind, auf Ihre eigene Situation anwenden oder sie entsprechend der für Sie relevanten Probleme umformulieren, ganz gleich, ob es dabei um Gesundheit, Sicherheit, Bildung, Geld, Berufsleben, Unterhaltung oder etwas anderes geht. Noise als Problem besser zu verstehen und Lösungsansätze dafür zu finden, ist Work in Progress und geht uns alle an. Wir alle können einen Beitrag dazu leisten. Dieses Buch wurde in der Hoffnung geschrieben, dass wir diese Chance auch tatsächlich nutzen.

# TEIL I
# Noise entdecken

Es ist nicht hinnehmbar, wenn straffällig gewordene Menschen für genau die gleiche Straftat unter ansonsten gleichen Bedingungen völlig unterschiedliche Strafmaße erhalten – zum Beispiel der eine eine Freiheitsstrafe von fünf Jahren und der andere Bewährung. Und doch passiert in vielen Ländern genau dies. Selbstverständlich ist das Strafjustizsystem auch von Bias durchdrungen. Aber konzentrieren wir uns in Kapitel 1 auf Noise – und vor allem auf das, was geschah, als ein berühmter Richter auf dieses Phänomen aufmerksam machte und eine Kampagne startete, die in gewissem Sinne die Welt veränderte (wenn auch nicht genug). Wir berichten hier über die Situation in den Vereinigten Staaten, aber wir sind überzeugt davon, dass es sich in vielen anderen Ländern ähnlich verhält. Wir vermuten, dass das Problem der Zufallsstreuung von Urteilen in einigen dieser Länder sogar noch gravierender ist als in den USA. Am Beispiel der Strafzumessung wollen wir zeigen, dass Noise zu großen Ungerechtigkeiten führen kann.

Die enorme Schwankungsbreite hat bei Strafurteilen besonders dramatische Auswirkungen, aber wir befassen uns auch mit dem Privatsektor, wo manchmal ebenso viel auf dem Spiel steht. Dies verdeutlichen wir in Kapitel 2 anhand einer großen Versicherungsgesellschaft. Dort haben sogenannte Underwriter – Mitarbeiter, die für eine Versicherung Sonderrisiken analysieren und auf dieser Grundlage Angebote erstellen – die Aufgabe, die Höhe von Versicherungsprämien für potenzielle Kunden festzusetzen, während Schadensregulierer den Wert geltend gemachter Schadensforderungen beurteilen müssen. Vielleicht nehmen Sie an, dass es sich um einfache und mechanische Aufgaben

handelt und dass verschiedene Fachleute ungefähr auf die gleichen Geldbeträge kommen. Wir haben ein sorgfältig geplantes Experiment – ein Noise Audit – durchgeführt, um diese Annahme zu überprüfen. Die Ergebnisse überraschten uns; aber sie haben auch die Führungsspitze des Unternehmens verblüfft, ja geradezu schockiert. Wir fanden nämlich heraus, dass das bloße Ausmaß an Noise das Unternehmen eine Menge Geld kostet. Das Experiment verdeutlicht, dass Noise auch erheblichen wirtschaftlichen Schaden verursachen kann.

Beide Beispiele – Strafjustizsystem und Versicherungswirtschaft – stützen sich auf Studien mit zahlreichen Teilnehmern, die eine große Zahl von Urteilen fällten. Aber viele wichtige Entscheidungen sind »singulär«, sie wiederholen sich nicht: Wie soll man mit einer scheinbar einzigartigen geschäftlichen Chance verfahren? Soll man ein völlig neues Produkt auf den Markt bringen? Wie sieht die angemessene Reaktion auf eine Pandemie aus? Soll man jemanden einstellen, der nicht dem Standardprofil entspricht? Tritt Noise auch bei Entscheidungen über solch einzigartige Situationen auf? Die Vermutung liegt nahe, dass dies nicht der Fall ist. Schließlich ist Rauschen unerwünschte Variabilität, und wie kann es bei einmaligen Entscheidungen zu Schwankungen kommen? In Kapitel 3 versuchen wir, diese Frage zu beantworten. Die Entscheidung, die man trifft – selbst in einer scheinbar einzigartigen Situation –, ist nur eine aus einer ganzen Wolke von Möglichkeiten. Auch hier findet man eine Menge Störgeräusche.

Aus diesen drei Kapiteln lässt sich eine Erkenntnis ableiten, die sich in einem Satz zusammenfassen lässt und wie ein roter Faden durch das gesamte Buch zieht: *Wo Urteile getroffen werden, gibt es Noise – und zwar mehr, als man gemeinhin erwartet.* Beginnen wir damit, herauszufinden, wie viel genau.

# KAPITEL I

# Verbrechen und Bestrafung: Ein Lotteriespiel

Angenommen, jemand ist wegen einer Straftat verurteilt worden – Ladendiebstahl, Heroinbesitz, Körperverletzung oder bewaffneter Raubüberfall. Welches Strafmaß hat er zu erwarten?

Die Antwort sollte nicht davon abhängen, welchem Richter oder welcher Richterin der Fall zufälligerweise übertragen wurde, ob es draußen heiß oder kalt ist oder die heimische Mannschaft am Vortag gewonnen hat. Es wäre empörend, wenn drei Personen, die wegen der gleichen Straftat verurteilt wurden, völlig unterschiedliche Strafmaße erhielten: Bewährung für den einen, zwei Jahre Haft für den zweiten und zehn Jahre für den dritten. Und trotzdem war und ist dieser Missstand in vielen Ländern nach wie vor Realität.

In der ganzen Welt besaßen Richter lange Zeit einen sehr großen Ermessensspielraum bei der Festsetzung des Strafmaßes. In vielen Ländern haben Experten dieses richterliche Ermessen positiv bewertet und es als gerecht und human gepriesen. Sie betonten beharrlich, das Strafmaß solle auf der Grundlage vieler Faktoren festgesetzt werden; dabei komme es nicht nur auf die Straftat an sich an, sondern auch auf den jeweiligen Charakter des Angeklagten und die Umstände der Tat. »Individueller Zuschnitt des Strafmaßes« lautete das Gebot der Stunde. Wären Richter durch Regeln in ihrer Ermessensfreiheit eingeschränkt, würden Straftäter auf eine menschenunwürdige Art und Weise behandelt; sie würden nicht als einzigartige Individuen betrachtet, die einen Anspruch darauf hätten, dass man ihre besonderen Lebensumstände berücksichtige. Die Idee eines »rechtsstaatlichen Verfahrens« an sich erschien vielen als ein Aufruf, Richtern unbeschränktes Ermessen zuzugestehen.

In den 1970er-Jahren begann die allgemeine Begeisterung für das richterliche Ermessen aus einem einfachen Grund zu schwinden: Es gab Belege dafür, dass die Strafzumessung einer Lotterie gleicht. Im Jahr 1973 lenkte ein berühmter US-Richter, Marvin Frankel, die öffentliche Aufmerksamkeit auf das Problem. Schon bevor er Richter geworden war, hatte sich Frankel nachdrücklich für Redefreiheit und Menschenrechte eingesetzt. Er war einer der Gründer des Lawyers Committee for Human Rights, einer Menschenrechtsorganisation, die heute Human Rights First heißt.

Frankel konnte ungestüm sein. Und er war empört über die Glückslotterie im Strafjustizsystem. Sein Buch *Criminal Sentences: Law Without Order* (»Strafurteile: Recht ohne Ordnung«) beginnt mit einer klaren Darlegung des Problems:

> *Jemand, der von einem Bundesgericht wegen Bankraubs verurteilt wurde, konnte eine Freiheitsstrafe von maximal 25 Jahren erhalten. Das heißt konkret, sein Strafmaß konnte irgendwo zwischen 0 und 25 Jahren liegen. Und wo genau es letztlich lag, hing, wie ich bald erkannte, weniger von dem Fall oder dem jeweiligen Angeklagten als von dem jeweiligen Richter ab, das heißt von den Ansichten, Vorlieben und Vorurteilen des Richters. Derselbe Angeklagte konnte daher in derselben Strafsache, je nachdem, welchem Richter der Fall übertragen wurde, zu sehr unterschiedlichen Strafen verurteilt werden.*

Frankel legte zur Untermauerung seines Arguments keine statistische Analyse vor. Aber er schilderte eine Reihe eindringlicher Beispiele, die die ungerechtfertigte Ungleichbehandlung von Personen vor Augen führten, die in allen strafrechtlich relevanten Punkten übereinstimmten. Zwei Männer, bislang beide nicht straffällig, wurden wegen Einlösung gefälschter Schecks über 58,40 beziehungsweise 35,20 Dollar verurteilt: der erste zu 15 *Jahren*, der zweite zu 30 *Tagen*. Für einander ähnelnde Fälle von Unterschlagung wurde ein Mann zu einer Freiheitsstrafe von 117 *Tagen* und ein anderer zu einer von 20 *Jahren* verurteilt. Unter Verweis auf zahlreiche ähnliche Fälle beklagte Frankel das, was er die »beinahe völlig unkontrollierten und weitreichenden Befugnisse« von Bundesrichtern nannte,[6] die dazu führten, dass »tagtäglich willkürliche Grausamkeiten begangen werden«,[7] die er in einem Gemeinwesen, »das von Ge-

setzen, nicht von menschlicher Willkür regiert wird«,[8] für nicht hinnehmbar ansah.

Frankel forderte den US-Kongress auf, dieser »Diskriminierung« – wie er die willkürlichen Grausamkeiten nannte – ein Ende zu setzen. Darunter verstand er hauptsächlich Noise in der Form nicht begründbarer Unterschiede bei der Strafzumessung. Aber auch Bias in Form »rassischer« und sozioökonomischer Ungleichheiten war ihm ein Dorn im Auge. Um sowohl Zufallsstreuung als auch systematische Abweichung zu bekämpfen, forderte er, die Ungleichbehandlung von Angeklagten dürfe nur dann zulässig sein, wenn die Unterschiede »durch sachdienliche Tests gerechtfertigt werden, die sich mit hinlänglicher Objektivität formulieren und anwenden lassen, um sicherzustellen, dass die Ergebnisse mehr sind als idiosynkratische Ukasse einzelner Beamter, Richter oder sonstiger Personen«.[9] (Der Ausdruck »idiosynkratische Ukasse« hört sich ein bisschen esoterisch an; darunter verstand Frankel »persönliche Erlasse«.) Aber er plädierte, weit darüber hinausgehend, für eine Verringerung von Noise durch ein detailliertes »Profil beziehungsweise eine Checkliste von Faktoren, die, wenn möglich, eine Form numerischer oder anderweitiger objektiver Einstufung umfassen sollten«.[10]

Da er dies zu Beginn der 1970er-Jahre schrieb, ging er nicht ganz so weit, die »Verdrängung von Menschen durch Maschinen« zu verteidigen. Aber er kam dem doch erstaunlich nahe. Er war fest davon überzeugt, dass »eine rechtsstaatliche Ordnung einen Korpus unpersönlicher Regeln erfordert, die allgemein anwendbar und für Richter genauso bindend sind wie für alle anderen«. Er sprach sich ausdrücklich für die Nutzung von »Computern als einem Hilfsmittel für geordnetes Denken bei der Strafzumessung« aus.[11] Er empfahl auch die Einrichtung einer Kommission für die Strafzumessung.[12]

Frankels Buch wurde zu einem der einflussreichsten in der gesamten Geschichte des Strafrechts – nicht nur in den Vereinigten Staaten, sondern weltweit. Sein Werk hatte jedoch handwerkliche Schwächen. Es stellte der Strafjustiz ein verheerendes Zeugnis aus, aber es war »impressionistisch«, zeigte Momentaufnahmen. Um zu überprüfen, ob Noise tatsächlich ein Problem war, wurden unmittelbar im Anschluss an die Publikation des Buches mehrere Studien durchgeführt, die dem Ausmaß von Noise bei Strafurteilen auf den Grund gingen.

Eine frühe, groß angelegte Studie dieser Art, die von Frankel selbst geleitet

wurde, fand 1974 statt. Fünfzig Richter aus verschiedenen Bezirken wurden gebeten, Strafen für Angeklagte in hypothetischen Fällen festzusetzen. Diese waren in identischen Berichten zusammengefasst worden, die man ihnen vor der Strafzumessung zur Verfügung stellte. Der wichtigste Befund lautete, dass »fehlender Konsens die Norm«[13] war und die Schwankungsbreite der Strafmaße »verblüffend«.[14] Ein Heroindealer konnte je nach Richter zu einem bis zehn Jahren Freiheitsstrafe verurteilt werden.[15] Die Freiheitsstrafen für einen Bankräuber konnten zwischen 5 und 18 Jahren betragen.[16] In einem Erpressungsfall reichten die Strafen von sage und schreibe 20 Jahren Haft und einer Geldstrafe von 65.000 Dollar zu lediglich 3 Jahren Haft und keiner Geldstrafe.[17] Besonders erschreckend: In 16 von 20 Fällen bestand kein Einvernehmen darüber, ob ein Freiheitsentzug überhaupt angemessen war.

An diese Studie schlossen sich eine Reihe weiterer an, die alle ähnlich schockierende Ausmaße an Noise zutage förderten. Im Jahr 1977 zum Beispiel führten William Austin und Thomas Williams eine Befragung von 47 Richtern durch; sie baten sie, in den gleichen fünf Fällen, die jeweils vergleichsweise geringfügige Vergehen betrafen, Strafmaße festzusetzen.[18]

Sämtliche Beschreibungen der Fälle enthielten auch Zusammenfassungen der Informationen, die Richter üblicherweise bei der Strafzumessung berücksichtigen, unter anderem die Anklage, die Zeugenaussagen, die Vorstrafen des Angeklagten (soweit vorhanden), sein sozialer Hintergrund und Hinweise auf seine Persönlichkeit. Der wichtigste Befund der Studie war eine »erhebliche Streuung«. In einem Fall, in dem es zum Beispiel um Einbruch ging, reichten die empfohlenen Haftstrafen von fünf Jahren bis zu lediglich dreißig Tagen (neben einer Geldstrafe von 100 Dollar). In einem anderen Fall um den Besitz von Marihuana empfahlen einige Richter Gefängnisstrafen, während andere zu Bewährung rieten.

Eine weitaus größere Studie, die 1981 durchgeführt wurde, bezog 208 Bundesrichter ein, denen dieselben 16 hypothetischen Fälle vorgelegt wurden.[19] Ihre wichtigsten Ergebnisse waren verblüffend: »In nur 3 der 16 Fälle bestand einhelliges Einvernehmen darüber, eine Haftstrafe zu verhängen. Selbst wenn sich die meisten Richter einig waren, dass eine Freiheitsstrafe angemessen wäre, gab es große Unterschiede in der Dauer der von ihnen empfohlenen Haftstrafen. In einer Betrugssache, in der sich die mittlere Haftstrafe auf 8,5 Jahre belief, war die längste Freiheitsstrafe, die verhängt wurde, lebenslänglich. In einem

anderen Fall war die mittlere Haftstrafe 1,1 Jahre, aber die längste Freiheitsstrafe, die empfohlen wurde, betrug 15 Jahre.«

So aufschlussreich diese Studien, die streng kontrollierte Experimente beinhalteten, auch sind, so unterschätzen sie doch höchstwahrscheinlich das Ausmaß an Noise in der realen Welt der Strafjustiz. Echte Richter erhalten viel mehr Informationen als die Teilnehmer in den sorgfältig ausgearbeiteten Falldarstellungen dieser Experimente. Ein Teil der zusätzlichen Informationen ist natürlich relevant, aber es gibt auch zahlreiche Anhaltspunkte dafür, dass irrelevante Informationen in Form nebensächlicher und scheinbar zufälliger Faktoren sich erheblich auf das Ergebnis auswirken können. So fand man zum Beispiel heraus, dass Richter am Tagesanfang oder nach einer Essenspause eine Strafe eher zur Bewährung aussetzen als unmittelbar vor einer solchen Pause. Wenn Richter hungrig sind, urteilen sie strenger.[20]

Bei einer Studie, in deren Rahmen Tausende von Jugendgerichtsentscheidungen ausgewertet wurden, kam Folgendes heraus: Wenn die lokale Footballmannschaft am Wochenende ein Spiel verliert, urteilen die Richter montags (und in geringerem Maß auch den Rest der Woche) strenger.[21] Angeklagte von dunkler Hautfarbe sind die Hauptleidtragenden dieser erhöhten Strenge. Eine andere Studie wertete 1,5 Millionen Gerichtsentscheidungen aus drei Jahrzehnten aus, wobei in ähnlicher Weise herauskam, dass Richter an Tagen, die auf eine Niederlage der örtlichen Footballmannschaft folgten, härter urteilten als an Tagen nach einem Sieg.[22]

Die Auswertung von sechs Millionen Gerichtsentscheidungen, die über einen Zeitraum von zwölf Jahren in Frankreich ergingen, ergab, dass Angeklagte an ihrem Geburtstag nachsichtiger behandelt werden.[23] (Wir vermuten, dass Richter auch an ihren eigenen Geburtstagen milder urteilen, aber soweit wir wissen, ist diese Hypothese nicht überprüft worden.) Sogar etwas scheinbar so Belangloses wie die Außentemperatur kann Richter beeinflussen.[24] Bei der Analyse von 207.000 asylbehördlichen Entscheidungen über einen Zeitraum von vier Jahren wurde ein deutlicher Effekt durch tägliche Temperaturschwankungen festgestellt: Wenn es draußen heiß ist, haben Asylbewerber schlechtere Aussichten, anerkannt zu werden. Wenn man in seinem Heimatland aus politischen Gründen verfolgt wird und in einem anderen Land Asyl beantragt, sollte man also hoffen und vielleicht sogar dafür beten, dass die Anhörung an einem kühlen Tag stattfindet.

## Zufallsschwankungen bei der Strafzumessung verringern

In den 1970er-Jahren wurde Edward M. Kennedy, der Bruder des ermordeten Präsidenten John F. Kennedy und eines der einflussreichsten Mitglieder des US-Senats, auf die Argumente Frankels und die empirischen Befunde, die seine Argumente stützten, aufmerksam. Kennedy war entsetzt. Schon 1975 brachte er einen Gesetzentwurf ein, der auf eine Reform der Strafzumessung abzielte; daraus wurde allerdings nichts. Aber Kennedy ließ sich dadurch nicht beirren. Auf die Studienergebnisse verweisend, drängte er weiterhin Jahr für Jahr auf die Verabschiedung dieses Gesetzentwurfs. Im Jahr 1984 gelang es ihm schließlich. Der Kongress reagierte auf die Belege für die ungerechtfertigte Streuung der Strafmaße mit der Verabschiedung des Sentencing Reform Act.[25] Das neue Gesetz sollte die Störgeräusche im System verringern, indem es »den uneingeschränkten Ermessensspielraum reduziert, den das Gesetz jenen Richtern und Bewährungsbehörden verleiht, die für die Verhängung und Vollstreckung der Strafen zuständig sind«.[26]

Kongressabgeordnete verwiesen auf die »nicht zu rechtfertigende Spannbreite« der Strafmaße, wobei sie insbesondere Befunde hervorhoben, wonach im Großraum New York Strafmaße für identische Fälle von 3 bis zu 20 Jahren Haft reichen konnten.[27] Wie von Richter Frankel empfohlen, ordnete das Gesetz die Einsetzung der United States Sentencing Commission an, deren Hauptaufgabe klar war: der Erlass von Leitlinien für die Strafzumessung, die bindend sein und die Bandbreite der Strafmaße – den Strafrahmen – einschränken sollten.

Im Jahr darauf erließ die Commission diese Leitlinien, die im Allgemeinen auf den Durchschnittsstrafen für ähnliche Verbrechen in einer Auswertung von 10.000 abgeschlossenen Strafprozessen basierten. Der Richter am Obersten Gerichtshof der USA, Stephen Breyer, der bei den Ausschussberatungen eine wichtige Rolle gespielt hatte, verteidigte die Bezugnahme auf die Strafmaßpraxis der Vergangenheit, indem er auf die unüberbrückbaren Meinungsunterschiede innerhalb der Commission hinwies: »Warum haben sich die Ausschussmitglieder nicht einfach zusammengesetzt und dieses Problem in einer rationalen Weise geklärt, statt einfach nur Urteile aus der Vergangenheit heranzuziehen? Die kurze Antwort lautet: Das konnten wir nicht. Wir konn-

ten es nicht, weil es jede Menge gute Argumente gibt, die in entgegengesetzte Richtungen weisen … Versuchen Sie einmal, alle Straftaten, die es gibt, in eine Rangfolge der Strafwürdigkeit zu bringen … Sammeln Sie dann die Ergebnisse von Ihren Freunden und prüfen Sie, ob sie alle übereinstimmen. Ich sage Ihnen, das wird nicht so sein.«[28]

Gemäß den Leitlinien müssen Richter bei der Strafzumessung zwei Faktoren berücksichtigen: die Straftat und eventuelle Vorstrafen des Angeklagten. Die Straftaten werden je nach ihrer Schwere in eine von 43 Stufen der »Strafwürdigkeit« eingeordnet. Das Vorstrafenregister des Angeklagten gibt Auskunft über die Anzahl und Schwere seiner früheren Verurteilungen. Sobald die aktuelle Straftat und die Vorstrafen zusammengeführt wurden, stecken die Richtlinien einen relativ engen Strafrahmen ab, wobei die höchstmögliche Strafe die geringstmögliche um sechs Monate oder 25 Prozent – je nachdem, was größer ist – übertreffen darf. Richtern ist es erlaubt, vom Strafrahmen abzuweichen, wenn sie strafverschärfende oder mildernde Umstände erkennen, aber die Abweichungen müssen gegenüber einem Berufungsgericht begründet werden.

Obgleich die Richtlinien bindend sind, sind sie nicht vollkommen starr. Sie gehen nicht annähernd so weit, wie es Richter Frankel wünschte. Sie geben Richtern noch immer erhebliche Entscheidungsspielräume. Trotzdem gelangten verschiedene Studien, die zahlreiche Methoden anwandten und unterschiedliche historische Zeiträume betrachteten, zu demselben Schluss: Die Richtlinien mindern Noise. Technischer ausgedrückt: Sie »reduzieren den Teil der Nettovariation des Strafmaßes, der auf den zufälligen Umstand der Identität des Richters, der das Urteil spricht, zurückzuführen ist«.[29]

Die aufwendigste Studie führte die Commission selbst durch.[30] Sie verglich Strafmaße bei Bankraub, Heroinhandel und Veruntreuung durch Bankmitarbeiter im Jahr 1985, vor dem Inkrafttreten der Leitlinien, mit den Strafmaßen, die zwischen dem 19. Januar 1989 und dem 30. September 1990 verhängt wurden. Straftäter wurden im Hinblick auf die Faktoren, die laut den Leitlinien für die Strafzumessung relevant waren, einander »passgenau zugeordnet«. Bei jeder Straftat waren die Unterschiede zwischen den Richtern im letztgenannten Zeitraum, nach der Verabschiedung des Sentencing Reform Act, viel geringer.

Laut einer anderen Studie betrug der durchschnittliche Unterschied in

der Dauer der von Richtern verhängten Strafen 17 Prozent beziehungsweise 4,9 Monate im Zeitraum 1986/87. Diese Zahl fiel zwischen 1988 und 1993 auf 11 Prozent oder 3,9 Monate.[31] Eine unabhängige Studie, die andere Zeiträume betrachtete, stieß auf ähnliche Erfolge bei der Verringerung von Disparitäten zwischen Richtern, definiert als die Unterschiede in den durchschnittlichen Strafmaßen, die Richter mit ähnlichen Fallbelastungen festsetzten.[32]

Ungeachtet dieser Befunde wurden die Leitlinien heftig kritisiert. Einige Personen, darunter viele Richter, hielten manche Strafmaße für zu hoch – was jedoch mit Bias, nicht mit Noise zu tun hat. Ein für unsere Zwecke viel interessanterer Einwand, der von zahlreichen Richtern erhoben wurde, lautete: Die Leitlinien seien zutiefst ungerecht, weil sie es Richtern untersagten, die besonderen Umstände der Tat angemessen zu berücksichtigen. Die Noise-Reduktion forderte einen Preis: Die Entscheidungsfindung wurde zu einem unannehmbar mechanischen Prozess. Die Juraprofessorin Kate Stith von der Universität Yale und Bundesrichter José Cabranes schrieben: »Das, was wir brauchen, ist nicht Blindheit, sondern Verständnis, Gerechtigkeit im Einzelfall« (im Sinne von Billigkeit), die »es nur in einem Urteil geben kann, das den Komplexitäten des Einzelfalls Rechnung trägt«.[33]

Diese Kritik führte dazu, dass die Leitlinien aus unterschiedlichen – teils juristischen, teils politischen – Gründen infrage gestellt wurden. Aber die Ablehnung blieb so lange folgenlos, bis der Oberste Gerichtshof aus technischen Gründen, die nichts mit der Debatte, die wir hier zusammenfassen, zu tun haben, die Leitlinien im Jahr 2005 für ungültig erklärte.[34] Aufgrund der Gerichtsentscheidung wurden sie zu bloßen unverbindlichen Entscheidungshilfen. Insbesondere die meisten Bundesrichter freuten sich sehr über dieses Urteil. 75 Prozent zogen fakultative Entscheidungshilfen vor, während nur 3 Prozent verbindlichen Regeln den Vorzug gaben.[35]

Welche Folgen hatte die Umwandlung der Leitlinien in unverbindliche Entscheidungshilfen? Die in Harvard lehrende Juraprofessorin Crystal Yang ging dieser Frage auf den Grund, nicht mit einem Experiment oder einer Umfrage, sondern mit einem riesigen Datensatz realer Strafmaße von fast 400.000 Verurteilten. Dabei fand sie insbesondere heraus, dass die Unterschiede zwischen Richtern nach 2005 deutlich zugenommen hatten. Als die Leitlinien verbindlich gewesen waren, hatten Angeklagte, die von einem relativ strengen Richter verurteilt worden waren, eine um 2,8 Monate längere Freiheitsstrafe

erhalten, als wenn sie von einem durchschnittlichen Richter verurteilt worden wären. Als die Richtlinien nur noch Entscheidungshilfen waren, verdoppelten sich die Unterschiede. Yang, die sich ganz ähnlich anhört wie Richter Frankel vierzig Jahre zuvor, schreibt, dass ihre »Befunde erhebliche Zweifel an der Gerechtigkeit von Strafurteilen wecken, da die Identität des Strafrichters, dem die Sache übertragen wird, in erheblichem Maße zu der Ungleichbehandlung ähnlicher Straftäter, die ähnlicher Straftaten schuldig gesprochen werden, beiträgt«.[36]

Nachdem die Leitlinien nur noch den Charakter von Empfehlungen hatten, urteilten Richter häufiger auf der Grundlage ihrer persönlichen Werte. Bindende Leitlinien reduzieren Bias und Noise. Nach der Entscheidung des Obersten Gerichtshofs gab es eine deutliche Zunahme der Unterschiede in den Strafmaßen afroamerikanischer und weißer Personen, die wegen der gleichen Straftat verurteilt wurden. Gleichzeitig nutzten Richterinnen ihren erweiterten Ermessensspielraum häufiger für ein milderes Urteil als Richter. Das Gleiche gilt für Richter, die von demokratischen Präsidenten ernannt wurden.

Drei Jahre nach dem Tod Richter Frankels im Jahr 2002 führte die Herabstufung der Leitlinien zu bloßen Entscheidungshilfen zum Rückfall in einen Zustand, der seinen Albträumen glich: Recht ohne Ordnung.

Die Geschichte des Kampfes von Richter Frankel für verbindliche Richtlinien der Strafzumessung vermittelt einen flüchtigen Eindruck von einigen der Schlüsselaspekte, die wir in diesem Buch behandeln werden.

Erstens: Die Urteilsbildung ist schwierig, weil die Welt ein komplexer, von Ungewissheiten geprägter Ort ist. Diese Komplexität wird offensichtlich in der Rechtsprechung, und sie trifft auch auf die meisten anderen Situationen zu, die ein »fachkundiges Urteil« erfordern. Dazu gehören Urteile, die von Ärzten, Pflegekräften, Juristen, Ingenieuren, Lehrern, Architekten, Hollywoodproduzenten, Mitgliedern von Berufungsausschüssen, Verlegern, Topmanagern aller Art und Mannschaftstrainern getroffen werden. Uneinigkeit ist unvermeidlich, wenn es um Urteile geht.

Zweitens: Das Ausmaß dieser Uneinigkeit ist viel größer, als wir erwarten. Während nur wenige Menschen den Grundsatz des richterlichen Ermessens ablehnen, stößt das Ausmaß der Disparität, die er erzeugt, auf fast einhellige Missbilligung. »System-Noise«, also die unerwünschte Uneinheitlichkeit von

Urteilen, die idealerweise vollkommen gleich sein sollten, kann zunehmende Ungerechtigkeit, hohe ökonomische Kosten und alle möglichen Arten von Fehlern verursachen.

Drittens: Noise lässt sich vermindern. Die von Richter Frankel befürwortete und von der US Sentencing Commission umgesetzte Strategie – Regeln und Richtlinien – ist eine von mehreren Methoden, die Noise wirksam reduzieren. Für andere Arten von Urteilen sind andere Vorgehensweisen besser geeignet. Einige Methoden zur Noise-Reduktion können gleichzeitig auch Bias verringern.

Viertens: Bemühungen um Noise-Reduktion stoßen oftmals auf Bedenken und erhebliche Widerstände. Auch diese müssen ausgeräumt werden, oder der Kampf gegen Noise wird scheitern.

## Zum Thema: Noise bei Strafurteilen

»Experimente zeigen große Unterschiede in den Strafmaßen, die verschiedene Richter in identischen Fällen festsetzen. Diese Schwankungen können nicht gerecht sein. Das Strafmaß eines Verurteilten sollte nicht davon abhängen, welchem Richter die Sache zufälligerweise zugewiesen wird.«

»Strafmaße sollten nicht von der Stimmung des Richters während der Verhandlung oder von der Außentemperatur abhängen.«

»Leitlinien sind eine Methode, um diesem Problem abzuhelfen. Aber viele Menschen mögen sie nicht, weil sie das richterliche Ermessen einschränken, das erforderlich sein kann, um Fairness und Genauigkeit zu gewährleisten. Schließlich ist doch jeder Fall einzigartig, oder?«

# KAPITEL 2
# Ein verrauschtes System

Unsere erste Begegnung mit Noise und das, was unser Interesse an dem Thema weckte, war nicht annähernd so dramatisch wie eine Berührung mit dem Strafjustizsystem. Tatsächlich war es eine Art Zufall und hatte mit einer Versicherungsgesellschaft zu tun, die das Beratungsunternehmen engagiert hatte, mit dem zwei von uns in Verbindung standen.

Das Thema Versicherungen ist nicht jedermanns Sache. Aber unsere Befunde verdeutlichen das Ausmaß des Noise-Problems in einer gewinnorientierten Organisation, die durch von Zufallsstreuung verfälschte Entscheidungen sehr viel verlieren kann. Unsere Erfahrungen mit der Versicherungsgesellschaft helfen zu verstehen, warum dieses Problem so oft unerkannt bleibt, und zeigen auf, was dagegen getan werden könnte.

Die Führungskräfte der Versicherung analysierten den potenziellen Nutzen von Maßnahmen, mit denen die Konsistenz der Urteile von Mitarbeitern verbessert werden sollte, die für das Unternehmen bedeutende finanzielle Entscheidungen trafen. Es ging der Chefetage also darum, Noise zu reduzieren. Alle waren sich einig, dass ein höheres Maß an Übereinstimmung wünschenswert wäre. Alle waren sich aber auch einig, dass Urteile niemals vollkommen deckungsgleich sein können, weil sie informell und teilweise subjektiv sind. Ein bisschen störendes Rauschen ist unvermeidlich.

Uneinigkeit herrschte in Bezug auf das Ausmaß von Noise: Die Führungskräfte bezweifelten, dass es ein gewichtiges Problem für ihr Unternehmen sei. Allerdings muss man es ihnen hoch anrechnen, dass sie sich bereit erklärten, die Frage mithilfe einer Art einfachem Experiment zu beantworten, das wir

bereits in der Einleitung kurz vorgestellt haben: das Noise Audit, mit dem gemessen werden soll, in welchem Ausmaß Fachkräfte voneinander abweichen, wenn sie im Rahmen einer Organisation die gleichen Aufgaben erfüllen. Das Ergebnis überraschte die Führungskräfte der Versicherung. Und es zeigte sich auch, dass damit das Noise-Problem auf perfekte Weise veranschaulicht wurde.

## Eine Lotterie, die Noise erzeugt

Viele Fachkräfte in Großunternehmen sind befugt, Entscheidungen zu treffen, die das Unternehmen binden. Die Versicherungsgesellschaft beschäftigt zum Beispiel eine erhebliche Anzahl von Underwritern, die Prämien für die Absicherung finanzieller Risiken festsetzen, etwa um eine Bank gegen Verluste aufgrund von Betrug oder sogenanntem Rogue-Trading (bei dem ein Mitarbeiter eigenmächtig nicht autorisierte Wertpapiergeschäfte tätigt) zu versichern. Die Gesellschaft beschäftigt auch viele Schadensregulierer, die die Kosten zukünftiger Schadensfälle prognostizieren und darüber verhandeln, falls und sobald sie eintreten.

In jeder größeren Niederlassung der Gesellschaft arbeiten mehrere qualifizierte Underwriter. Wenn von einem Kunden ein Preisangebot erbeten wird, kann jedem, der zufälligerweise gerade verfügbar ist, die Aufgabe übertragen werden, es zu erstellen. Tatsächlich wird der konkrete Underwriter, der das Angebot erarbeitet, durch eine Lotterie ausgewählt.

Der genaue Wert des Angebots hat erhebliche Folgen für das Unternehmen. Eine hohe Prämie ist vorteilhaft, wenn das Angebot angenommen wird, jedoch mit dem Risiko verbunden, den Kunden an einen Wettbewerber zu verlieren. Eine niedrige Prämie wird eher akzeptiert, man verzichtet jedoch auf Einnahmen, die möglicherweise benötigt werden, um einen Verlust abzudecken. Für jedes Risiko gibt es einen – um eine Metapher aus der Wirtschaftswelt zu verwenden – »Goldlöckchen«-Preis, das heißt einen, der genau richtig, weder zu hoch noch zu niedrig ist. Und es ist recht wahrscheinlich, dass das durchschnittliche Urteil einer großen Zahl von Fachleuten nicht allzu weit von dieser *Goldilocks*-Zahl entfernt ist. Preise, die höher oder niedriger sind als diese Zahl, sind kostspielig – dies erklärt, wieso die Streuung verrauschter Urteile die Ertragskraft von Unternehmen schmälert.

Die Tätigkeit von Schadensregulierern wirkt sich ebenfalls auf die Finanzen der Versicherung aus, für die sie arbeiten. Nehmen wir zum Beispiel an, ein Schaden wird im Namen eines Arbeitnehmers (des Schadensberechtigten) gemeldet, der nach einem Arbeitsunfall seine rechte Hand dauerhaft nicht mehr gebrauchen kann. Der Schadensfall wird einem Regulierer übertragen – weil er oder sie ebenfalls gerade verfügbar ist. Der Regulierer sammelt sämtliche Fakten des Falls und erstellt eine Schätzung der Gesamtkosten, die auf den Versicherer zukommen. Dann übernimmt es der Regulierer, mit dem Vertreter des Schadensberechtigten zu verhandeln, um sicherzustellen, dass dieser die in der Versicherungspolice versprochenen Leistungen erhält, während er selbst zugleich das Unternehmen davor schützen soll, überhöhte Zahlungen zu leisten.

Die frühe Schätzung des Regulierers spielt eine wichtige Rolle, weil sie bei zukünftigen Verhandlungen mit dem Anspruchsberechtigten eine implizite Zielgröße ist. Die Versicherungsgesellschaft ist auch rechtlich dazu verpflichtet, Rückstellungen in Höhe der prognostizierten Kosten für jeden Schadensfall zu bilden (also genug Geld auf die Seite zu legen, um den Geschädigten abzufinden). Auch hier gibt es aus Sicht der Gesellschaft einen »Goldlöckchen«-Wert. Eine erfolgreiche außergerichtliche Schadensregulierung ist nicht garantiert, weil der Schadensberechtigte einen Rechtsbeistand hat, der vielleicht klagt, wenn ihm das Angebot als unzureichend erscheint. Andererseits könnte eine übermäßig großzügige Rückstellung dem Regulierer einen zu großen Spielraum einräumen, um ungerechtfertigten Forderungen nachzugeben. Sein Urteil hat für die Gesellschaft – und mehr noch für den Schadensberechtigten – weitreichende Folgen.

Unsere Wahl des Wortes »Lotterie« unterstreicht hier die Rolle des Zufalls bei der Auswahl eines Underwriters oder Regulierers. Normalerweise wird ein Versicherungsfall einer einzelnen Fachkraft zugewiesen, und niemand weiß, was geschehen wäre, wenn ein anderer Kollege ausgewählt worden wäre.

Lotterien dieser Art können sinnvoll und müssen nicht ungerecht sein. Mithilfe gesellschaftlich akzeptierter Lotterien werden »positive Dinge« wie etwa die Teilnehmerplätze für Lehrveranstaltungen an einer Universität zugeteilt oder auch »negative« wie die Einberufung zum Militär. Sie dienen einem Zweck. Aber die Urteilslotterien, über die wir sprechen, teilen nichts zu. Sie produzieren lediglich Ungewissheit. Stellen wir uns eine Versicherungsgesellschaft vor, deren Underwriter frei von Noise entscheiden und alle die optimale

Prämie festsetzen. Aber dann greift ein Zufallsgenerator ein, um das Angebot, das dem Kunden vorgelegt wird, abzuändern. Für eine solche Lotterie gäbe es offensichtlich keine Rechtfertigung. Und es gibt auch keine Rechtfertigung für ein System, in dem das Ergebnis von der Identität der konkreten Person abhängt, die zufällig ausgewählt wurde, um ein fachkundiges Urteil zu fällen.

## Ein Noise Audit entdeckt System-Noise

Die Lotterie, die einen bestimmten Richter auswählt, um ein Strafurteil zu sprechen, oder einen einzelnen Schützen als Stellvertreter seiner Mannschaft, erzeugt Variabilität, aber diese Variabilität bleibt unbemerkt. Ein Noise Audit – so wie es bei US-Bundesrichtern bezüglich der Strafzumessung durchgeführt wurde (vgl. Kapitel 6) – ist eine Methode, um sie aufzuspüren. Bei einem solchen Audit wird derselbe Fall von zahlreichen Personen bewertet, und die Schwankungsbreite ihrer Antworten wird sichtbar gemacht.

Die Urteile von Underwritern und Schadensregulierern eignen sich besonders gut für diese Übung, weil ihre Entscheidungen auf schriftlichen Informationen beruhen. Zur Vorbereitung auf das Noise Audit bei der Versicherungsgesellschaft verfassten Führungskräfte ausführliche Beschreibungen von fünf repräsentativen Fällen für jede der beiden Gruppen – Underwriter und Schadensregulierer. Dann wurden die Mitarbeiter gebeten, jeweils unabhängig voneinander zwei oder drei Fälle zu beurteilen. Es wurde ihnen nicht gesagt, dass der Zweck der Studie darin bestand, die Streuung ihrer Beurteilungen zu erfassen.[37]

Bevor Sie weiterlesen, sollten Sie kurz darüber nachdenken, was Sie selbst auf die folgende Frage antworten würden: »Wenn Sie in einer gut geführten Versicherungsgesellschaft zwei sachkundige Underwriter oder Schadensregulierer nach dem Zufallsprinzip auswählen würden, wie weit lägen deren Schätzungen für denselben Fall Ihres Erachtens wohl auseinander?« Konkret gefragt: Wie groß wäre die Differenz zwischen beiden Schätzungen in Prozent ihres Durchschnittswerts?

Wir baten nun zahlreiche Führungskräfte der Versicherungsgesellschaft um ihre Antwort auf diese Frage, und in den darauffolgenden Jahren erhielten wir Schätzungen eines breiten Spektrums von Personen, die in unterschied-

lichen Berufen im Unternehmen tätig waren. Erstaunlicherweise stach eine Antwort aus allen anderen hervor. Die meisten Führungskräfte der Versicherung schätzten 10 Prozent oder weniger. Als wir auch 828 CEOs und Topmanager von Unternehmen aus den unterschiedlichsten Branchen fragten, welche Unterschiede sie bei ähnlichen fachkundigen Urteilen erwarteten, war »10 Prozent« die mittlere Antwort und zugleich die häufigste (die zweithäufigste war »15 Prozent«). Eine Differenz von 10 Prozent würde zum Beispiel bedeuten, dass einer der beiden Underwriter in demselben Fall eine Prämie von 9.500 Dollar festsetzt, während der andere eine Prämie von 10.500 Dollar anbietet. Keine unerhebliche Differenz, aber doch eine, die eine Organisation wohl tolerieren kann.

Unser Noise Audit enthüllte jedoch viel größere Unterschiede. Nach Auswertung unserer Daten gelangten wir zu dem Ergebnis, dass die mittlere Differenz beim Underwriting 55 Prozent betrug, also etwa das Fünffache dessen, was die meisten Leute, einschließlich der Führungskräfte der Versicherung, erwarten würden. Dieser Befund bedeutet zum Beispiel: Wenn ein Underwriter eine Prämie von 9.500 Dollar festsetzt, lautet das Angebot des zweiten nicht etwa 10.500, sondern 16.700 Dollar. Für Schadensregulierer betrug der Median der Differenz 43 Prozent. Wir betonen, dass diese Ergebnisse Mediane, also Zentralwerte sind, das heißt, bei der Hälfte der Fallpaare war die Differenz zwischen den beiden Prämien sogar noch größer.

Die Führungskräfte, denen wir von den Ergebnissen des Noise Audits berichteten, erkannten schnell, dass das bloße Ausmaß an Noise ein kostspieliges Problem darstellte. Ein Topmanager meinte sogar, die jährlichen Kosten, die Noise im Underwriting unter Berücksichtigung sowohl der Einbußen im Neugeschäft aufgrund überhöhter Angebote als auch der Verluste infolge zu niedrig tarifierter Verträge verursache, beliefen sich auf Hunderte Millionen Dollar.

Niemand konnte genau sagen, wie groß der Fehler (oder das Bias) war, weil niemand mit Sicherheit den »Goldlöckchen«-Wert für jeden Fall kannte. Aber man musste nicht die Mitte der Zielscheiben sehen, um die Streuung auf ihrer Rückseite zu messen und zu erkennen, dass diese Uneinheitlichkeit ein Problem war. Die Daten zeigten, dass der Preis, der von einem Kunden verlangt wird, in einem fragwürdigen Ausmaß von der Lotterie abhängt, die den Mitarbeiter auswählt, der den Geschäftsvorgang bearbeitet. Gelinde gesagt, würde es

Kunden nicht gefallen, zu erfahren, dass sie ohne ihr Einverständnis bei einer solchen Lotterie mitmachen müssen.

Ganz allgemein erwarten die Menschen im Umgang mit privaten und öffentlichen Organisationen ein System, das in verlässlicher Weise konsistente Entscheidungen trifft. Sie erwarten darin keine Störgeräusche, kein System-Noise.

## Unerwünschte Streuung im Gegensatz zu erwünschter Vielfalt

Ein definierendes Merkmal von System-Noise ist, dass das Rauschen »unerwünscht« ist, aber wir sollten gleich hier betonen, dass die Streuung von Urteilen nicht immer unerwünscht ist.

Betrachten wir Dinge, bei denen es um Präferenzen oder Geschmacksvorlieben geht. Wenn sich zehn Filmkritiker denselben Film ansehen, wenn zehn Weinverkoster denselben Wein beurteilen oder wenn zehn Bücherwürmer denselben Roman lesen, dann erwarten wir nicht, dass sie die gleiche Meinung haben. Die Vielfalt der geschmacklichen Vorlieben ist erwünscht und entspricht vollkommen unseren Erwartungen. Niemand möchte in einer Welt leben, in der jeder genau die gleichen Vorlieben und Abneigungen hat. (Nun ja, fast niemand.) Allerdings kann die Vielfalt von Vorlieben helfen, Fehler zu erklären – nämlich dann, wenn eine persönliche Präferenz mit einem fachkundigen Urteil verwechselt wird. Wenn eine Filmproduzentin beschließt, ein ungewöhnliches Projekt zu realisieren (zum Beispiel über den Aufstieg und Niedergang des Telefons mit Wählscheibe), weil ihr persönlich das Drehbuch gefällt, begeht sie vielleicht einen großen Fehler, wenn es sonst niemandem zusagt.

Auch in einer Wettbewerbssituation, in der die besten Urteile belohnt werden, wird erwartet und begrüßt, wenn es eine Bandbreite von Urteilen gibt. Wenn mehrere Unternehmen oder mehrere Teams in derselben Organisation miteinander konkurrieren, um innovative Lösungen für dasselbe Kundenproblem zu finden, wollen wir nicht, dass sie alle dieselbe Vorgehensweise anwenden. Das Gleiche gilt, wenn mehrere Forschergruppen ein wissenschaftliches Problem, wie etwa die Entwicklung eines Impfstoffs, in Angriff nehmen: Wir wollen durchaus, dass sie es aus verschiedenen Perspektiven betrachten. Selbst

Prognostiker verhalten sich manchmal wie Spieler in einem Wettkampf. Der Analyst, der eine Rezession, die sonst niemand erwartete, zutreffend vorhergesagt hat, kann sich sicher sein, berühmt zu werden, während derjenige, der sich nie vom Konsens entfernt, unbekannt bleibt. Auch in solchen Situationen wird die Vielfalt von Ideen und Urteilen begrüßt, weil Streuung nur der erste Schritt ist. In einer zweiten Phase werden die Ergebnisse dieser Urteile nämlich gegeneinander ins Rennen geschickt, und das beste wird sich durchsetzen. Auf einem Markt kann Selektion ohne Variation ebenso wenig funktionieren wie in der Natur.

Geschmacksfragen und Wettbewerbssituationen werfen interessante Urteilsprobleme auf. Aber wir konzentrieren uns hier auf Urteile, bei denen Variabilität *unerwünscht* ist. System-Noise ist ein Problem von »Systemen« – das heißt von Organisationen, nicht von Märkten. Wenn Börsenhändler den Wert einer Aktie unterschiedlich beurteilen, verdienen einige von ihnen Geld, während andere dies nicht tun. Märkte beruhen auf unterschiedlichen Einschätzungen. Aber wenn einer dieser Händler zufällig ausgewählt würde, um diese Einschätzung im Namen seiner Firma vorzunehmen, und wenn wir herausfänden, dass seine Kollegen zu völlig anderen Einschätzungen gelangten, dann gäbe es System-Noise in dieser Firma, und das wäre ein Problem.

Als wir unsere Ergebnisse den Topmanagern einer Vermögensverwaltungsgesellschaft vorstellten, war dies für sie ein Ansporn, ein eigenes Noise Audit durchzuführen, das dieses Problem auf elegante Weise veranschaulicht. Sie baten 42 erfahrene Anlageexperten des Unternehmens darum, den fairen Wert einer Aktie abzuschätzen, sprich den Kurs, zu dem sie diese weder kaufen noch verkaufen würden. Die Investoren stützten ihre Analyse auf eine einseitige Beschreibung des betreffenden Unternehmens; die Daten umfassten eine vereinfachte Gewinn- und Verlustrechnung, die Bilanz und Kapitalflussrechnungen der vergangenen drei sowie Projektionen für die nächsten zwei Jahre. Der Noise-Medianwert, der auf die gleiche Weise wie bei der Versicherungsgesellschaft ermittelt wurde, betrug 41 Prozent. Derart große Differenzen unter Anlageexperten im selben Unternehmen, die die gleichen Bewertungsmethoden anwenden, sind keine gute Nachricht.

Wo immer die Person, die eine Beurteilung vornimmt, aus einer Gesamtheit gleich gut qualifizierter Individuen zufällig ausgewählt wird – wie es bei dieser Vermögensverwaltungsgesellschaft, im Strafjustizsystem und in der frü-

her diskutierten Versicherungsgesellschaft der Fall ist –, stellt Noise ein Problem dar. Viele Organisationen schlagen sich mit System-Noise herum: Ein Zuweisungsverfahren, das tatsächlich zufallsabhängig ist, bestimmt oftmals, von welchem Arzt man in einem Krankenhaus behandelt wird, welcher Richter einen Fall verhandelt, welcher Patentprüfer einen Antrag bearbeitet, welcher Kundendienstmitarbeiter eine Beschwerde entgegennimmt und so weiter. Unerwünschte Variabilität bei den dabei getroffenen Urteilen kann gravierende Probleme verursachen, wie etwa Geldverlust und grassierende Ungerechtigkeit.

Häufig wird irrigerweise angenommen, die unerwünschte Streuung von Urteilen sei unerheblich, weil sich Zufallsabweichungen angeblich »gegenseitig aufheben«. Es stimmt, dass sich positive und negative Abweichungen bei einem Urteil über denselben Sachverhalt tendenziell gegenseitig aufheben, und wir werden noch ausführlich diskutieren, wie man sich diese Eigenschaft zunutze machen kann, um Noise zu reduzieren. Aber mit Noise behaftete Systeme treffen nicht mehrere Urteile über denselben Sachverhalt. Sie fällen vielmehr verrauschte Urteile über verschiedene Sachverhalte. Wenn eine Versicherungspolice überteuert und eine andere zu billig ist, dann mag die Preisgestaltung »im Durchschnitt« richtig erscheinen, aber die Versicherungsgesellschaft hat zwei kostspielige Fehler gemacht. Wenn zwei Straftäter, die beide zu Freiheitsstrafen von fünf Jahren verurteilt werden sollten, einmal drei Jahre und einmal sieben Jahre bekommen, ist das kein »im Durchschnitt« gerechtes Ergebnis. In Systemen, die von Noise betroffen sind, heben sich Fehler nicht gegenseitig auf. Vielmehr summieren sie sich.

## Die Illusion der Übereinstimmung

Zahlreiche Publikationen, die zum Teil mehrere Jahrzehnte zurückreichen, haben die Existenz von Noise in fachlichen Urteilen dokumentiert. Da wir diese Arbeiten kannten, haben uns die Ergebnisse des Noise Audits bei der Versicherungsgesellschaft nicht überrascht. Überrascht hat uns allerdings die Reaktion der Führungskräfte, denen wir unsere Ergebnisse präsentierten: Niemand im Unternehmen hatte auch nur annähernd das von uns beobachtete Ausmaß an Noise erwartet. Niemand stellte die Aussagekraft des Audits infrage, und niemand behauptete, dass das beobachtete Ausmaß an Noise akzep-

tabel sei. Doch hatte es den Anschein, dass das Problem – und seine hohen Kosten – für das Unternehmen neu war. Noise war wie eine undichte Stelle im Keller. Es wurde toleriert, nicht weil man es für akzeptabel hielt, sondern weil es unbemerkt geblieben war.

Wie war das möglich? Wie konnten Fachkräfte mit der gleichen Aufgabe und im selben Büro so weit auseinanderliegen, ohne dies zu bemerken? Wie konnten Führungskräfte diese Tatsache übersehen, in der sie nun eine ernsthafte Gefahr für die Leistungsfähigkeit und Reputation ihres Unternehmens erkannten? Uns wurde klar, dass das Problem des System-Noise in Organisationen oft unerkannt bleibt und die weitverbreitete Taubheit gegenüber den Störgeräuschen genauso interessant ist wie ihre Häufigkeit. Die Noise Audits deuteten darauf hin, dass anerkannte Fachleute – und die Organisationen, die sie beschäftigen – eine *Illusion der Übereinstimmung* aufrechterhielten, während sie tatsächlich in ihren täglichen fachlichen Urteilen nicht übereinstimmten.

Um zu verstehen, wie diese Illusion entsteht, sollten Sie sich in einen Underwriter an einem normalen Arbeitstag hineinversetzen. Sie haben mehr als fünf Jahre Berufserfahrung, Sie wissen, dass Sie von Ihren Kollegen geschätzt werden, und Sie respektieren und mögen sie. Sie wissen, dass Sie Ihre Arbeit gut machen. Nachdem Sie die komplexen Risiken, denen ein Finanzunternehmen ausgesetzt ist, analysiert haben, gelangen Sie zu dem Schluss, dass eine Prämie von 200.000 Dollar angemessen ist. Das Problem ist komplex, unterscheidet sich aber nicht erheblich von denjenigen, mit denen Sie sich Tag für Tag befassen.

Stellen Sie sich jetzt vor, dass Ihre Kollegen im Büro die gleichen Informationen erhalten und das gleiche Risiko bewerten. Könnten Sie glauben, dass mindestens die Hälfte von ihnen oder mehr eine Prämie festgesetzt hat, die entweder höher ist als 255.000 oder niedriger als 145.000 Dollar? Unwahrscheinlich. Tatsächlich nehmen wir an, dass Underwriter, die von dem Noise Audit hörten und es auch für aussagekräftig hielten, niemals wirklich glaubten, dass die Schlussfolgerungen daraus auf sie persönlich zutrafen.

Die meisten von uns leben die meiste Zeit mit der nicht hinterfragten Überzeugung, dass »wir die Welt genauso wahrnehmen, wie sie nun einmal ist«. Es ist nur ein kleiner Schritt von dieser Überzeugung zu jener: »Andere Menschen nehmen die Welt genauso wahr, wie ich es tue.« Diese Anschauung, der sogenannte *naive Realismus*, ist von zentraler Bedeutung für unsere Über-

zeugung, dass wir mit allen anderen Menschen eine bestimmte Sicht der Wirklichkeit teilen. Wir stellen diese Überzeugung nur selten infrage.[38] Zu jedem beliebigen Zeitpunkt interpretieren wir die Welt um uns herum nach einem einzigen Modell, und wir bemühen uns normalerweise nicht ernsthaft darum, plausible Alternativen zu finden. Eine Interpretation ist ausreichend, und wir erleben sie als wahr. Wir malen uns keine alternativen Sichtweisen der Wirklichkeit aus.

Im Falle fachlicher Urteile wird die Überzeugung, dass andere die Welt ganz ähnlich sehen wie wir, Tag für Tag auf vielfältige Weise bestätigt. Erstens teilen wir mit unseren Kollegen eine gemeinsame Sprache und eine Reihe von Regeln über die Gesichtspunkte, die bei unseren Entscheidungen berücksichtigt werden sollten. Wir stimmen mit anderen auch darin überein, dass Urteile, die gegen diese Regeln verstoßen, wertlos sind. Die gelegentlichen Meinungsverschiedenheiten mit Kollegen erleben wir als Fehlurteile ihrerseits. Wir haben kaum Gelegenheit, zu bemerken, dass unsere vereinbarten Regeln vage sind – sie reichen aus, um einige Möglichkeiten auszuschließen, aber nicht, um eine gemeinsame positive Beurteilung eines bestimmten Falls sicherzustellen. Wir können reibungslos mit Kollegen zusammenarbeiten, ohne jemals zu bemerken, dass sie die Welt nicht so sehen, wie wir es tun.

Eine von uns interviewte Underwriterin beschrieb, wie sie durch Erfahrung in ihrer Abteilung ein wachsendes Gefühl der Sicherheit entwickelt hatte: »Als ich neu war, besprach ich 75 Prozent der Fälle mit meinem Vorgesetzten … Nach ein paar Jahren war das nicht mehr nötig – ich gelte jetzt als Expertin … Im Lauf der Zeit habe ich immer mehr auf mein Urteilsvermögen vertraut.« Wie viele von uns entwickelte auch sie hauptsächlich durch praktische Übung Vertrauen in ihr Urteilsvermögen.

Wir verstehen die Psychologie dieses Prozesses recht gut. Selbstvertrauen wird durch die subjektive Erfahrung von Urteilen gestärkt, die mit zunehmender Geläufigkeit und Leichtigkeit getroffen werden, zum Teil weil sie Urteilen gleichen, die in ähnlichen Fällen in der Vergangenheit getroffen wurden. In dem Maße, wie die Underwriterin im Lauf der Zeit lernte, mit ihrem früheren Selbst übereinzustimmen, wuchs ihr Vertrauen in ihre Urteilskraft. Nichts deutete darauf hin, dass sie – nach der anfänglichen Lehrphase – gelernt hatte, eine einheitliche Linie mit anderen zu finden, überprüft hatte, in welchem Ausmaß sie mit anderen übereinstimmte, oder sich überhaupt bemühte, zu ver-

hindern, dass ihre Vorgehensweisen immer stärker von denen ihrer Kollegen abwichen.

Im Fall der Versicherungsgesellschaft wurde die Illusion der Übereinstimmung erst durch das Noise Audit erschüttert. Wieso haben die Führungskräfte des Unternehmens ihr Noise-Problem nicht bemerkt? Es gibt mehrere mögliche Antworten auf diese Frage, aber eine, die in zahlreichen Situationen eine große Rolle zu spielen scheint, ist schlichtweg das mit Uneinigkeit verbundene Unbehagen. Die meisten Organisationen schätzen Konsens und Eintracht mehr als Widerspruch und Konflikt. Die bestehenden Abläufe scheinen oftmals ausdrücklich dafür ausgelegt zu sein, die Häufigkeit, mit der Mitarbeiter anderen Meinungen ausgesetzt werden, zu minimieren, und, wenn solche Uneinigkeiten auftreten, diese wegzuerklären.

Nathan Kuncel, Professor für Psychologie an der Universität von Minnesota und ein führender Forscher auf dem Gebiet der Leistungsvorhersage, berichtete uns von einem Fall, der dieses Problem verdeutlicht. Kuncel half einer Schulzulassungsstelle, ihren Entscheidungsprozess zu optimieren. Zuerst las eine Person einen Aufnahmeantrag, beurteilte diesen und leitete ihn mit der Beurteilung an einen zweiten Gutachter weiter, der ihn seinerseits beurteilte. Kuncel empfahl – aus Gründen, die in diesem Buch noch deutlich werden –, die Beurteilung des ersten Gutachters zu verdecken, damit sie den zweiten nicht beeinflusse. Die Antwort der Schule lautete: »Wir haben das früher so gemacht, aber es führte zu so vielen Konflikten, dass wir auf das gegenwärtige System umstellten.« Diese Schule ist nicht die einzige Organisation, die Konfliktvermeidung für mindestens genauso wichtig erachtet wie das Treffen der richtigen Entscheidung.

Betrachten wir einen weiteren Mechanismus, auf den viele Unternehmen zurückgreifen: nachträgliche Analysen von Fehlentscheidungen. Als Lernmethode sind diese sogenannten Post-mortem-Analysen nützlich. Aber wenn wirklich ein Fehler gemacht wurde, in dem Sinne, dass ein Urteil weit von professionellen Normen abwich, ist es nicht schwierig, darüber zu diskutieren. Experten werden unschwer zu dem Schluss gelangen, dass es weit vom Konsens entfernt ist. (Sie werden es vielleicht auch als seltene Ausnahme abtun.) Fehlurteile lassen sich viel leichter identifizieren als richtige Urteile. Das Anprangern eklatanter Fehler und die Ausgrenzung leistungsschwacher Kollegen wird Fachkräften nicht helfen, zu erkennen, wie uneinig sie sind, wenn sie

im Großen und Ganzen annehmbare Urteile treffen. Im Gegenteil, der leicht erzielbare Konsens über schlechte Urteile wird die Illusion der Übereinstimmung vielleicht sogar verstärken. Die eigentliche Lektion – die Allgegenwart von System-Noise – werden sie dann nicht lernen.

Wir hoffen, dass Sie allmählich beginnen, in System-Noise ein ernsthaftes Problem zu erkennen. Seine Existenz ist keine Überraschung. Noise ist eine Folge der informellen Natur der Urteilsbildung. Doch wie wir in diesem Buch immer wieder sehen werden, wird das Ausmaß des Noise, das zu beobachten ist, wenn eine Organisation genau hinsieht, fast immer als schockierend erlebt. Unser Fazit ist einfach: *Überall, wo Urteile getroffen werden, gibt es Noise, und zwar mehr, als man denkt.*

## Zum Thema: System-Noise in einer Versicherungsgesellschaft

»Wir sind abhängig von der Qualität fachkundiger Urteile, von Underwritern, Schadensregulierern und anderen. Wir weisen jeden Fall einem Experten zu, aber wir gehen von der falschen Annahme aus, dass ein anderer Experte zu einem ähnlichen Urteil kommt.«

»Es gibt fünfmal mehr System-Noise, als wir dachten – beziehungsweise als wir tolerieren können. Ohne ein Noise Audit hätten wir dies nie erkannt. Das Noise Audit hat die Illusion der Übereinstimmung zerstört.«

»System-Noise ist ein ernstes Problem: Es kostet uns Hunderte Millionen Dollar.«

»Überall, wo Urteile getroffen werden, gibt es Noise – und mehr davon, als man denkt.«

# KAPITEL 3
# Einmalige Entscheidungen

Die Fallstudien, die wir bislang erörtert haben, beziehen sich auf Entscheidungen, die wiederholt getroffen werden. Was ist das richtige Strafmaß für jemanden, der wegen Diebstahls verurteilt wurde? Was ist die richtige Prämie für ein bestimmtes Risiko? Obwohl jeder Fall in einem gewissen Sinne einzigartig ist, sind solche Urteile *wiederkehrende Entscheidungen*. Ärzte, die Diagnosen stellen, Richter, die Bewährungsfälle verhandeln, Mitarbeiter von Zulassungsstellen, die Anträge prüfen, Steuerberater, die Steuerunterlagen ausfüllen – dies alles sind Beispiele für wiederkehrende Entscheidungen.

Rauschen in wiederkehrenden Entscheidungen wird durch ein Noise Audit nachgewiesen, wie wir es im letzten Kapitel vorgestellt haben. Unerwünschte Streuung lässt sich leicht definieren und messen, wenn untereinander austauschbare Fachkräfte in ähnlichen Fällen Entscheidungen treffen. Viel schwieriger oder vielleicht sogar unmöglich ist es, die Idee des Noise auf eine Kategorie von Urteilen anzuwenden, die wir *einmalige Entscheidungen* nennen.

Betrachten wir zum Beispiel die Krise, mit der die Welt im Jahr 2014 konfrontiert war. In Westafrika starben zahlreiche Menschen an Ebola. Aufgrund der engen weltweiten Verflechtungen ließen Prognosen befürchten, dass sich die Infektionskrankheit rasch über die gesamte Welt ausbreiten und Europa und Nordamerika besonders hart treffen würde. In den Vereinigten Staaten wurden Forderungen laut, den Flugverkehr aus den betroffenen Regionen einzustellen und die Grenzen zu schließen. Der politische Druck, entsprechende Maßnahmen zu ergreifen, wuchs, und bekannte und gut informierte Persönlichkeiten sprachen sich für diese Schritte aus.

Präsident Obama sah sich mit einer der schwierigsten Entscheidungen seiner Amtszeit konfrontiert – einer, die sich ihm bislang noch nicht gestellt hatte und der er sich auch kein weiteres Mal gegenübersehen würde. Er beschloss, die Grenzen nicht zu schließen. Stattdessen schickte er 3.000 Helfer – medizinische Fachkräfte und Soldaten – nach Westafrika. Er führte eine bunt gemischte internationale Koalition von Staaten an, die nicht immer gut zusammenarbeitete; er nutzte ihre Ressourcen und ihren Sachverstand, um das Problem an der Wurzel anzupacken.

## Einmalig im Gegensatz zu wiederkehrend

Entscheidungen, die nur ein einziges Mal getroffen werden, wie Präsident Obamas Reaktion auf die Ebola-Epidemie, sind »einmalig«, weil sie nicht wiederholt von derselben Person oder Gruppe getroffen werden, es keine vorgefertigte Standardantwort gibt und sie wirklich einzigartige Merkmale aufweisen. Beim Umgang mit der Ebola-Epidemie konnte sich Präsident Obama auf keine wirklichen Präzedenzfälle stützen. Wichtige politische Entscheidungen sind – ebenso wie die schicksalsträchtigsten militärischen – oftmals gute Beispiele für einmalige Entscheidungen.

Im privaten Bereich haben Entscheidungen, die man bei der Auswahl eines Arbeitsplatzes, beim Kauf eines Hauses oder bei Heiratsanträgen trifft, die gleichen Merkmale. Selbst wenn dies nicht Ihre erste Stelle, Ihr erstes Haus oder Ihre erste Ehe ist, und ungeachtet der Tatsache, dass zahllose Menschen vor Ihnen mit dieser Entscheidung konfrontiert waren, fühlt sie sich für Sie einzigartig an. Im Geschäftsleben müssen Firmenchefs oft Entscheidungen treffen, die ihnen selbst einzigartig erscheinen: Ob eine potenziell bahnbrechende Innovation am Markt eingeführt werden soll, wie lange der Betrieb während einer Pandemie geschlossen bleiben soll, ob eine Geschäftsstelle in einem anderen Land eröffnet werden soll oder ob man behördliche Auflagen einfach hinnehmen soll.

Zwischen einmaligen und wiederkehrenden Entscheidungen besteht kein kategorialer Unterschied, vielmehr liegen sie auf einem Kontinuum. Underwriter haben möglicherweise mit einigen Fällen zu tun, die ihnen sehr ungewöhnlich vorkommen. Umgekehrt gilt: Wenn Sie zum vierten Mal in Ihrem Leben

ein Haus kaufen, ist für Sie der Hauskauf zu einer wiederkehrenden Entscheidung geworden. Aber Extrembeispiele deuten darauf hin, dass der Unterschied bedeutsam ist. Es ist eine Sache, in den Krieg zu ziehen, eine andere, jährliche Budget-Revisionsberichte durchzulesen.

## Noise bei einmaligen Entscheidungen

Herkömmlicherweise wird klar differenziert zwischen einmaligen Entscheidungen und wiederkehrenden Beurteilungen, die austauschbare Mitarbeiter in großen Organisationen regelmäßig vornehmen. Während wiederkehrende Entscheidungen von Sozialwissenschaftlern erforscht werden, sind einmalige, folgenreiche Entscheidungen die Domäne von Historikern und Managementgurus. Die wissenschaftlichen Herangehensweisen an die beiden Arten von Entscheidungen sind recht unterschiedlich. Bei der Analyse wiederkehrender Entscheidungen werden oft statistische Methoden angewandt; Sozialwissenschaftler werten damit viele ähnliche Entscheidungen aus, um Muster zu erkennen, Regelmäßigkeiten zu identifizieren und die Genauigkeit zu messen. Diskussionen einmaliger Entscheidungen dagegen verfolgen in der Regel einen kausalen Ansatz; sie betrachten die Ereignisse in der Rückschau und konzentrieren sich darauf, die Ursachen dafür zu ermitteln. Historische Analysen, aber auch Fallstudien über erfolgreiche oder gescheiterte Führungskräfte zielen darauf ab, zu verstehen, wie eine im Grunde einmalige Entscheidung getroffen wurde.

Die Natur einmaliger Entscheidungen wirft eine wichtige Frage für die Erforschung von Noise auf. Wir haben es definiert als unerwünschte Streuung von Urteilen über das gleiche Problem. Da einmalige Probleme niemals in völlig identischer Form erneut auftreten, trifft diese Definition nicht auf sie zu. Schließlich ist der Gang der Geschichte einzigartig. Die Entscheidung Obamas im Jahr 2014, medizinisches Fachpersonal und Soldaten nach Westafrika zu schicken, lässt sich nicht mit den Entscheidungen anderer amerikanischer Präsidenten in Bezug auf dieses spezifische Problem zu diesem konkreten Zeitpunkt vergleichen (auch wenn man darüber spekulieren kann). Vielleicht sind Sie damit einverstanden, Ihre Entscheidung, diese konkrete Person zu heiraten, mit den Entscheidungen anderer Menschen, die Ihnen ähneln, zu vergle-

chen, aber dieser Vergleich wird für Sie nicht so bedeutungsvoll sein wie der Vergleich, den wir zwischen den Angeboten von Underwritern in Bezug auf denselben Fall angestellt haben. Sie und Ihr Ehepartner sind einzigartig. Die Anwesenheit von Noise in einmaligen Entscheidungen lässt sich nicht auf direktem Wege beobachten.

Aber einmalige Entscheidungen sind auch nicht frei von den Faktoren, die bei wiederkehrenden Entscheidungen Noise erzeugen. Auf dem Schießstand benutzen die Schützen von Team C (das mit dem verrauschten Ergebnis) womöglich Gewehre mit Zielfernrohren, die alle nicht sauber justiert sind, oder vielleicht haben sie auch einfach nur zittrige Hände. Wenn wir nur den ersten Schützen beobachten würden, hätten wir keine Vorstellung davon, wie sehr das Team von Noise beeinflusst ist, aber die Quellen des Rauschens wären trotzdem da. In ähnlicher Weise müssen Sie sich, wenn Sie eine einmalige Entscheidung treffen, vorstellen, dass ein anderer Entscheider, selbst wenn er genauso kompetent wäre wie Sie und die gleichen Ziele und Werte hätte, aus den gleichen Fakten nicht genau den gleichen Schluss ziehen würde. Und als Entscheider sollten Sie zugeben, dass Sie vielleicht eine andere Entscheidung getroffen hätten, wenn einige vermeintlich unerhebliche Aspekte der Situation oder des Entscheidungsprozesses anders gewesen wären.

Kurzum, bei einer einmaligen Entscheidung können wir Noise nicht messen, aber wenn wir kontrafaktisch denken, wissen wir mit Sicherheit, dass Noise vorhanden ist. So wie die unruhige Hand des Schützen bedeutet, dass ein einzelner Schuss irgendwo anders hätte landen *können*, so bedeutet Noise bei Entscheidern und im Entscheidungsprozess, dass die einmalige Entscheidung anders hätte ausfallen *können*.

Betrachten wir all die Faktoren, die sich auf eine einmalige Entscheidung auswirken. Wenn die Experten, die für die Analyse der Ebola-Bedrohung und die Erstellung von Reaktionsplänen zuständig waren, andere Personen gewesen wären, mit anderen Hintergründen und Lebenserfahrungen, wäre ihr Vorschlag an Präsident Obama der gleiche gewesen? Wenn dieselben Fakten auf eine leicht andere Weise präsentiert worden wären, wäre die Diskussion dann genauso verlaufen? Wenn die wichtigsten Akteure in einer anderen Stimmung gewesen wären oder sich während eines Schneesturms getroffen hätten, wäre die endgültige Entscheidung dann eine andere gewesen? So betrachtet, erscheint die einmalige Entscheidung nicht mehr determiniert zu sein. In Ab-