Subjektive Informationstheorie

Verminderung von Risiko durch Sachinformationsfragen

Jochen Koubek
Technische Hochschule Darmstadt
Schlossgartenstr.7

10. Dezember 1995

Inhaltsverzeichnis

Einleitung **i**

1 Grundlagen **1**
1.1 Unsicherheit . . . 1
1.2 Logische und Arithmetische Operationen . . . 3
1.3 Subjektive Wahrscheinlichkeit . . . 5
1.3.1 Einschätzung von Zufallsgrößen . . . 6
1.3.2 Plausibilitäten . . . 9
1.4 Bedingte Plausibilitäten . . . 10
1.5 Stochastische Unabhängigkeit . . . 12
1.6 Verteilungen . . . 13
1.6.1 Endliche Verteilungen . . . 13
1.6.2 Verteilungen auf [a,b] . . . 13
1.7 Axiomatik . . . 14

2 Sachinformationsfragen und Information **15**
2.1 Heuristik . . . 15
2.2 Information . . . 19
2.3 Eigenschaften des Informationsmaßes . . . 20
2.4 Eindeutigkeit des Informationsmaßes . . . 24
2.5 Bedingte Fragen . . . 27
2.6 Transinformation . . . 29
2.7 Antworten . . . 31
2.8 Fragestrategien . . . 32
2.8.1 Exkurs: Bäume . . . 32

3 Die Sätze von Shannon **34**
3.1 Quellen . . . 34
3.1.1 Gedächtnislose Quellen . . . 35
3.1.2 Markov-Quellen . . . 37
3.2 Kanäle . . . 39
3.3 Codes . . . 40
3.3.1 Eindeutig decodierbare Codes . . . 40
3.3.2 Der 1. Satz von Shannon . . . 41
3.4 Der gestörte Kanal . . . 46
3.4.1 Die Sätze von Shannon . . . 48
3.5 Entscheidungsprobleme . . . 49

4 Information und Entscheidung **52**
4.1 Heuristik . . . 52
4.1.1 Exkurs: Logik . . . 54
4.1.2 Algorithmus . . . 55

4.2 Entscheidungstheorie 57
4.2.1 Das Entscheidungsproblem 58
4.2.2 Entscheidungskriterien 60
4.2.3 Der Wert einer Frage 63
4.2.4 Kostengünstige Fragen 66
4.2.5 Entscheidung bei Risiko 67
4.2.6 Sicherheitsäquivalente 69
4.2.7 Axiomatik des Bernoulli-Prinzips 71

5 Abschließende Bemerkungen **74**

Denn so ist das Leben. Und anders.

Einleitung

In dieser Arbeit habe ich versucht, Informationstheorie, wie sie vor fünfzig Jahren von dem amerikanischen Nachrichtentechniker Claude Shannon eingeführt worden ist, aus subjektivistischer Sicht zu schildern und mit der Entscheidungstheorie zu verbinden. Die hier vorgestellten Konzepte sind ein Modell für eine Entscheidungssituation, in dem das entscheidende Subjekt in einer Welt handeln muß, die sich in einem unbekannten Zustand befindet. Die Menge möglicher Zustände, die in der Entscheidungssituation als verschieden aufgefaßt werden, ist endlich. Jede Handlung führt in Verbindung mit einem dieser Zustände zu einem Ergebnis, das von dem Entscheidungsträger im voraus absehbar ist. Es fehlt lediglich das Wissen um den tatsächlich vorliegenden Weltzustand. Dieses Nicht-Wissen wird aufgefaßt als Desinformation, die durch geeignete Fragen reduziert werden soll. Information und Desinformation werden hierbei als subjektive Größe betrachtet und spiegeln so jeweils nur das Wissen oder Nicht-Wissen eines Subjektes wider. Die ganze Arbeit gliedert sich in fünf Kapitel. Leitidee ist die Vorstellung von Information als Verminderung von Unsicherheit. Diese Unsicherheit wird durch Fragen formuliert.
Im *ersten Kapitel* wird Unsicherheit eingeführt mit Hilfe der Begriffe der Wahrscheinlichkeitstheorie. Die subjektive Unsicherheit über das Eintreten eines Ereignisses wird ausgedrückt durch subjektive Wahrscheinlichkeiten oder Plausibilitäten. Diese Plausibilitäten werden gemessen durch die Bereitschaft, auf das Eintreten eines bestimmtes Ereignisses zu wetten. Ausgehend von dem Begriff der Plausibilität werden andere Grundbegriffe, wie bedingte Plausibilität oder Verteilung ebenfalls aus subjektivistischer Sicht eingeführt. Im *zweiten Kapitel* geht es darum, Information als Antworten auf Sachinformationsfragen zu interpretieren. Sachinformationsfragen sind Fragen, auf die dem Subjekt eine endliche Anzahl möglicher Antworten bekannt sind und lediglich unklar ist, welche davon zutrifft. Jede dieser Antworten wird mit einer bestimmten Plausibilität belegt, zusätzlich trifft genau eine Antwort zu. Eine Frage ist somit eine Menge von Antworten zuzüglich eines Plausibilitätsmaßes auf diesen Antworten. Danach wird ein Maß für die Information angegeben, die beim Beantworten einer solchen Frage gewonnen wird. Eine konkret gegebene Antwort muß aber nicht die Frage vollständig beantworten, es können immer noch Unsicherheiten bestehen bleiben. Daher wird eine Antwort interpretiert als eine Transformation einer Frage in eine andere Frage. Im Grenzfall ist dies die leere Frage, d.h. eine Frage, die gar keine Information mehr verlangt. Eine solche Frage ist beantwortet. Das *dritte Kapitel* bettet die Ergebnisse der klassische Informationstheorie

in die bis dahin eingeführten Begrifflichkeiten ein und zeigt, daß das Übertragen von Zeichen über einen Kanal als Frage-und-Antwort Prozeß interpretiert werden kann. Die hier gestellte Frage an die Quelle lautet immer „Welches ist das nächste ausgegebene Zeichen ?“. Die Vernachlässigung jeglicher Semantik wird in dieser Frage sehr deutlich. Die Antwort wird über den Kanal gesendet, in Form einer Zeichenkette. Codes sind hier Fragefolgen, wenn das Alphabet, das zur Übertragung genutzt wird, kleiner ist, als das Alphabet der angeschlossenen Quelle. Binäre Codes teilen das Quellalphabet in zwei möglichst gleichplausible Gruppen und es wird zunächst die Zugehörigkeit des gesendeten Zeichens zu einer dieser Gruppen erfragt, diese wird wiederum unterteilt, usw. Im Falle des gestörten Kanals wird die Ausgangsfrage nicht vollständig beantwortet, es bleibt Informationsbedarf. Schwerpunkt dieses Kapitels sind natürlich die Sätze von Shannon, die im wesentlichen aussagen, daß unter bestimmten Bedingungen selbst bei einem beliebig gestörten Kanal ein Code gefunden werden kann, der eine beliebig genaue Übertragung ermöglicht. Die Übertragungsgeschwindigkeit liegt hierbei beliebig dicht an der Produktionsgeschwindigkeit der Quelle. Das *vierte Kapitel* beschäftigt sich schließlich mit Entscheidungen bei Risiko. Risiko ist hierbei abhängig vom Grad der Desinformation. Eine Entscheidungssituation läßt sich beschreiben als eine Menge möglicher Handlungen und eine Menge an Zuständen, in denen sich die Umwelt befinden kann. Eine Handlung führt in Verbindung mit einem Umweltzustand zu einem Ergebnis, verbunden mit einem bestimmten Nutzen. Das Problem besteht darin, handeln zu müssen, ohne den tatsächlichen Umweltzustand zu kennen. Jedem Umweltzustand wird von dem Subjekt eine bestimmte Plausibilität zugeordnet. Die klassische Entscheidungstheorie beschränkt sich darauf, aus diesen Daten eine rationale Handlung abzuleiten. Als rationales Kriterium gilt hier das Bernoulli-Kriterium, das Handlungen nach dem Erwartungswert ihres Nutzens bewertet. Demgegenüber wird hier noch die Möglichkeit aufgezeigt, über den Umweltzustand Fragen zu stellen, um die Plauibilitäten zu verschieben und das Entscheidungsrisiko zu vermindern. Günstigstenfalls wird subjektive Sicherheit erzeugt, wenn die Frage „Welcher Umweltzustand liegt vor ?“ vollständig beantwortet wird. Die Verbindung von Entscheidungstheorie und Informationstheorie ist der eigentlich originelle Teil dieser Arbeit und wird meines Erachtens in der Literatur bis zum gegenwärtigen Zeitpunkt noch nicht behandelt. Das *fünfte Kapitel* faßt noch einmal die Ergebnisse axiomatisch zusammen. Axiomatisiert werden die Logik der Sicherheit als klassische Prädikatenlogik, die Logik der Unsicherheit als Plausibilitätskalkül, die Messung von Information durch die Shannonsche Formel und das Bernoulli-Kriterium als rationale Entscheidungsregel. Diese Axiome können dank ihrer unmittelbaren Einsichtigkeit angesehen werden als Grundlagen rationalen Verhaltens. Aber ebenso, wie die Gesetze der Logik einer Argumentation nicht zugrunde gelegt werden müssen, müssen diese Axiome nicht zur Handlungsgrundlage gemacht werden. Allerdings würde eine solche Argumentation ebenso wie eine solche Handlung als irrational eingestuft. Das Ergebnis dieser Arbeit ist ein Modell. Modellierung ist immer auch Reduktion des Weltganzen

auf ausgewählte Teilaspekte. Der wichtigste Aspekt ist hier der der Rationalität, ich gehe von dem fiktiven *homo oeconomicus* aus, der rational seinen Nutzen zu optimieren versucht. Bei aller Eleganz der Ergebnisse möchte ich aber nicht unerwähnt lassen, daß dies nur *ein* Aspekt dieser Welt ist, die anderen hier allerdings wenig Berücksichtigung gefunden haben. Ebenso gehen in die Modellierung zahlreiche persönliche Sichtweisen ein, die zwar dem einen offensichtlich erscheinen mögen, für den anderen aber bloße Konstrukte sind. Ich erhebe nicht den Anspruch, *die* Welt beschrieben, sondern einen Ausschnitt meiner Weltsicht in die Sprache Mathematik übersetzt zu haben.

Danken möchte ich meinem Betreuer Herrn Prof.Krabs, der es nicht nur zugelassen, sondern mich auch dazu ermuntert hat, meine Arbeit in die Richtung zu lenken, die mich am meisten interessierte. Ebenfalls danke ich allen, die sich meinen Enthusiasmus über das Thema anhören wollten oder mußten, so daß viele Gedanken erst durch diese Diskussionen gereift sind. Stellvertretend danke ich hier Antje Jürgens und Carsten Tibke, die sich zusätzlich die Mühe gemacht haben, die Arbeit auf inhaltliche und orthographische Konsistenz zu überprüfen.

Für Cécile

1 Grundlagen

„Wahrscheinlichkeit gibt es nicht“. So beginnt das Buch “Teoria Delle Probabilità“, im Englischen “Theory of Probability“ von Bruno de Finetti, [Df]. In diesem Buch wird der Begriff der Wahrscheinlichkeit aus subjektivistischer Sicht geschildert und diesen Ansatz möchte ich auch zur Grundlage der vorliegenden Arbeit machen. In diesem ersten Kapitel geht es nicht nur darum, wichtige Begriffe aus der Wahrscheinlichkeitstheorie vorzustellen, sondern eben auch diese Begriffe aus der Perspektive eines erkennenden und handelnden Subjektes zu entwickeln. Dieser Ansatz steht im gewissen Gegensatz zu den maßtheoretischen Fundierungen, deren populärste Form die Axiome von Kolmogoroff sind. Der wesentliche Unterschied ist die Verneinung der σ-Additivität des Mengensystems der Ereignisse. Aber auch die Annahme, alle Ereignisse auf Basiszustände zurückführen zu können und so Ereignisse als Teilmengen eines Zustandsraumes aufzufassen, ist hier lediglich eine Veranschaulichung. Als Konsequenz ergibt sich, daß die Verteilungsgesetze nicht einzig und allein durch die Verteilungsfunktion festgelegt sind, sondern daß auch diese Funktion lediglich ein Hilfsmittel ist, mit dem sich Verteilungen unter anderem beschreiben lassen. Ich werde diese Diskussion aber in dieser Arbeit nicht fortsetzen und führe den Begriff der Verteilung durch den der Verteilungsfunktion ein.

1.1 Unsicherheit

In vielen Fällen muß ein Subjekt mit einer Welt umgehen, deren Zustand ihm unklar ist, im Falle des Wetters ebenso wie im Verhalten von Wählern, den klassischen Beispielen wie Münz- oder Würfelwurf oder subatomare Vorgänge. Diesen Fällen ist gemeinsam, daß die Umwelt sich in bestimmten Zuständen befinden kann, über deren Eintreten oder Vorliegen das Subjekt keine Sicherheit hat. Dennoch gehe ich davon aus, daß diese Situationen in sich wohl determiniert bzw. begründet sind, es liegt also am erkennenden Subjekt, das diese Gesetzmäßigkeiten und Begründungen nicht zu erkennen vermag[1]. Im folgenden werde ich dieses erkennende Subjekt mit S bezeichnen. Die Zustände, in denen die Umwelt sich nach Meinung von S befinden kann, werden unterschieden von dem Zustand, in dem sie sich tatsächlich befindet. Im Sinne der Wahrscheinlichkeitstheorie (W-Theorie) werde ich von *Ereignissen* sprechen, um die Zustände der Welt zu charakterisieren. Im Gegensatz zur klassischen W-Theorie wird hier nicht davon ausgegangen, daß es vergleichbare Grund-Zustände gibt, aus denen sich die Ereignisse zusammensetzen lassen. Zwei unterschiedliche Weltzustände sind tatsächlich unterschiedlich, auch wenn sie die gleichen Eigenschaften oder

[1] ich gehe nicht davon aus, daß menschlichem Verhalten irgendwelche (z.B. historische Gesetzmäßigkeiten) zugrunde liegen. Wenn ich in diesem Zusammenhang von “Begründung“ spreche, so meine ich, daß Menschen sich nie ohne Grund verhalten, auch wenn ihnen selber die Gründe nicht immer klar sind. Die Pointe ist eben, daß sie menschlicher Erkenntnis wohl nie ganz zugänglich sein werden.

(subjektiven) Wahrscheinlichkeiten zugeordnet haben. So trifft das Ereignis „es wird eine gerade Zahl gewürfelt“ genau dann ein, wenn entweder der Würfel-Zustand “2“, “4“ oder “6“ realisiert wird. Auch wenn der Wurf des Würfels von physikalischen Gesetzen und Parametern wie Wurfhöhe, -geschwindigkeit, -winkel, Luftströmungen, Bodenbeschaffenheit, usw. determiniert ist, besteht für S keine Möglichkeit, die Rechnungen durchzuführen, die nötig wären, um die gewürfelte Zahl zu ermitteln. Über das Eintreffen des Ereignisses „es wird eine gerade Zahl gewürfelt“ besteht also für S eine gewisse Unsicherheit. Ein erneutes Würfeln wäre ein anderes Ereignis, auch wenn die Unsicherheit die selbe bleibt. Die Ereignisse, die, abhängig von der jeweiligen Situation, eintreffen können oder nicht, werden mit ihren sprachlichen Beschreibungen gleichgesetzt. Das heißt, daß das Eintreffen eines Ereignisses E identisch ist mit dem Wahr-sein der Aussage „Ereignis E trifft ein“. Ich orientiere mich hierbei an Wittgensteins “Tractatus“[Wi], der beginnt mit den Sätzen: „1. Die Welt ist alles, was der Fall ist.“ „1.1 Die Welt ist die Ansammlung der Tatsachen, nicht der Dinge“. „2 Was der Fall ist, die Tatsache, ist das Bestehen von Sachverhalten“. Oder anders: ein Ereignis ist alles und nur das, was Inhalt einer Aussage ist. Die Menge der Ereignisse und die Menge möglicher Aussagesätze gelten für diese Arbeit als identisch. Ich will an dieser Stelle nicht eine metaphysische Diskussion über erkenntnistheoretische Fragen führen, weil dies eine Arbeit im Bereich der Mathematik ist. Dieses Abbrechen der Diskussion demonstriert sehr schön das Wesen von Philosophie als Disziplin, die all die Dinge in Frage stellt, die die anderen Disziplinen brauchen, um überhaupt arbeiten zu können. Ich setze also meine durchaus streitbare These über Ereignisse als Festsetzung an den Anfang und werde darauf aufbauen. Die Aussagen „Ereignis E tritt ein“ und „Die Aussage: „Ereignis E tritt ein“ ist wahr“ werden im folgenden synonym gebraucht. Ein Ereignis tritt ein, oder es tritt nicht ein. Diese beiden Zustände werde ich als 'wahr' und als 'falsch', bzw. mit den Zahlenwerten 1=wahr, das Ereignis tritt ein und 0=falsch, d.h. das Ereignis tritt nicht ein, bezeichnen. Wenn die Rede von Wahrscheinlichkeiten und später von Plausibilitäten ist, so heißt das, daß S die Ereignisse als nicht sicher betrachtet, weil nicht genügend Informationen vorliegen, oder weil eine Berechnung aus Komplexitätsgründen z.B. nicht möglich ist. Diese Wahrscheinlichkeit liegt zwischen der Sicherheit des Eintretens des Ereignisses, und der Sicherheit des Nicht-Eintretens, also zwischen 0 und 1. Ebenso kann Unsicherheit bestehen über den Wert, den eine Größe X hat. Dieser Wert kann z.B. aus dem Bereich der natürlichen oder reellen Zahlen kommen. Der wahre Wert ist eindeutig, dem Subjekt S aber nicht bekannt. In einem solchen Fall kann man von *Zufallsgröße* sprechen. Ein Ereignis ist also eine spezielle Zufallsgröße, eine, die nur die Werte 0 oder 1 annehmen kann. Die Grenzen, in denen sich der Wert der Zufallsgröße bewegen kann, bezeichne ich mit $\inf(X) \leq X \leq \sup(X)$. Der Begriff der Zufallsgröße orientiert sich hier also nicht an dem der Meßbarkeit bestimmter Mengen. Dieser wichtige Unterschied führt dazu, daß viele Konzepte nicht in der aus der maßtheoretisch orientierten Wahrscheinlichkeitstheorie bekannten Allgemeinheit und Eleganz eingeführt werden

können. Erkauft wird damit allerdings ein verständlicherer und ehrlicherer Umgang mit Unsicherheiten. Von allen möglichen Werten, die eine Zufallsgöße X annehmen kann, dem Alternativenraum $\mathcal{S}$, kommen für das Subjekt nur einige in Frage. Die Menge der (für S) möglichen Alternativen wird mit $\mathcal{Q}$ bezeichnet. Die Wahrscheinlichkeitsrechnung, wie sie hier vorgestellt wird, ist eine Logik der Unsicherheit, vgl. [Df], [Ra], [Sa]. Insofern müssen logische Operationen auf Ereignissen, bzw. auf den sie beschreibenden Aussagen, interpretiert werden.

1.2 Logische und Arithmetische Operationen

Durch die Festsetzung "wahr=1, falsch=0" läßt sich mit Ereignissen im Sinne boolscher Operationen rechnen, mit

$$\wedge \text{ logisches Produkt; } \vee \text{ logische Summe; } \neg \text{ Negation .}$$

Für Zahlen gelten die arithmetischen Operationen:

$$\cdot \text{ Produkt; } + \text{ Summe, mit den Inversen: / und -}$$

Für Zahlen gelten folgende Definitionen:

$$x \wedge y = \min(x, y),\ x \vee y = \max(x, y),\ \neg x = 1 - x$$

Diese Operationen werden allerdings nur im Zusammenhang mit den Zahlen 0 und 1, d.h. den Wahrheitswerten von Ereignissen genutzt. Für Zahlen, wie für Ereignisse gelten folgende Gleichungen:

$$\begin{aligned}
\neg(x \wedge y) &= \neg x \vee \neg y \\
\neg(x \vee y) &= \neg x \wedge \neg y \\
x \wedge (y \vee z) &= (x \wedge y) \vee (x \wedge z) \\
x \vee (y \wedge z) &= (x \vee y) \wedge (x \vee z) \\
x \wedge x &= x \\
x \vee x &= x
\end{aligned}$$

Das logische Produkt zweier Ereignisse A und B ist das Ereignis E, das genau dann wahr ist, wenn beide Faktoren wahr sind. Da beide Faktoren nur die Werte 0 oder 1 annehmen können, stimmt das logische Produkt mit dem arithmetischen überein und wird als $E = AB$ bezeichnet. Die Negation eines Ereignisses E ist wahr, wenn E falsch ist und umgekehrt. Es gilt also

$$\neg E = 1 - E$$

Die logische Summe ist wahr, wenn mindestens einer der Summanden wahr ist und stimmt mit der arithmetischen Summe überein. Mit den oben angeführten

Eigenschaften läßt sich die logische Summe auch wie folgt schreiben:

$$\begin{aligned} A \vee B &= \neg(\neg A \wedge \neg B) \\ &= 1-(1-A)(1-B) = A+B-AB. \\ &\quad \text{sowie} \\ A \vee B \vee C &= 1-(1-A)(1-B)(1-C) \\ &= A+B+C-AB-AC-BC+ABC \\ &\quad \text{und allgemein} \\ E_1 \vee E_2 \vee \ldots \vee E_n &= \sum_i E_i - \sum_{i<j} E_i E_j \\ &\quad + \sum_{i<j<h} E_i E_j E_h - \ldots \pm E_1 E_2 \cdots E_n \end{aligned}$$

Die arithmetische Summe zweier Ereignisse ist kein Ereignis, sondern eine Zufallszahl, die die Anzahl der Erfolge angibt.

$$\text{(logische Summe)} = 1 \wedge \text{(arithmetische Summe)}$$

Der Wahrheitsgehalt eines Ereignisses wird von dem Subjekt S festgestellt. Hiefür wird das Symbol $\vdash$ eingeführt. $\vdash E$ besagt also: Das Ereignis E ist sicher für eine Person. $\vdash \neg E$ heißt, daß E als unmöglich eingestuft wird, $\neg \vdash E$ heißt, daß E für S möglich ist. Die *Unvereinbarkeit* zweier Ereignisse bedeutet, daß es unmöglich ist, daß beide eintreffen: $\vdash \neg AB$. Zwei Ereignisse A und B heißen *vollständig*, wenn es unmöglich ist, daß beide zugleich nicht eintreffen: $\vdash \neg(\neg A \neg B)$. Diese Definitionen lassen sich leicht auf n Ereignisse verallgemeinern, sogar auf abzählbar viele. Mengentheoretisch bedeutet dies, daß eine höchstens abzählbare Familie von Mengen die Menge Q aller möglichen Punkte überdeckt. Eine *Einteilung* ist eine Familie vollständiger und unvereinbarer Ereignisse, d.h., es ist sicher, daß genau ein Ereignis eintrifft. Diese Familie kann endlich oder unendlich sein. Im wesentlichen werden in dieser Arbeit allerdings nur endliche Einteilungen betrachtet. Zwei Ereignisse A und B heißen *unabhängig*, wenn das Subjekt S über B trotz vollständiger Information über den Zustand von A nicht mehr weiß, als ohne das Wissen über A und umgekehrt. Ebenso heißen n Ereignisse *unabhängig*, wenn alle 2^n wahr/falsch-Kombinationen eintreffen können, d.h. daß jedes Ereignis selbst nach der Bestimmung der Werte aller übrigen Ereignisse unbestimmt bleibt. Sind n Ereignisse unabhängig, so auch alle ihre Teilmengen. Die Umkehrung ist i.allg. nicht richtig, selbst wenn alle echten Teilmengen unabhängig sind, kann die gesamte Menge abhängig sein. Ein Ereignis E ist genau dann *abhängig* von den Ereignissen $E_1, E_2, \ldots, E_n$, wenn es durch eine logische Funktion dieser Ereignisse ausdrückbar ist. Zufallsgrößen sind *unabhängig*, wenn unter keinen Umständen das Wissen um eine die Unsicherheit um die anderen verändert. Wenn X, Y, Z

jeweils r, s, t mögliche Werte annehmen können, dann sind alle rst Tripel mögliche Werte für (X, Y, Z), d.h. der Raum Q möglicher Werte für (X, Y, Z) ist das Cartesische Produkt der Mengen Q_X, Q_Y und Q_Z. Auf der anderen Seite ist z.B. Z abhängig von X und Y, wenn Z sich als Funktion $Z = f(X, Y)$ von X und Y ausdrücken läßt.

1.3 Subjektive Wahrscheinlichkeit

In Situationen der Unsicherheit, wie im vorigen Abschnitt eingeführt, hat S nun die verschiedenen Zustände beschrieben und zu Ereignissen zusammengefaßt. Dabei soll es aber nicht bleiben, die bestehende Unsicherheit soll quantifiziert werden. Unter allen Zuständen, die eintreten können, werden manche als vermutlich eher zutreffend angesehen als andere. Dieses Ansehen ist kein Voraussagen, wie es der Fall bei Wahrsagern oder Propheten oder unseriösen Zukunftsforschern ist. Diese versuchen, die bestehende Unsicherheit in irgendeine Sicherheit umzuwandeln, indem ein Zustand als bestimmt zutreffend prognostiziert wird. Dies soll hier nicht geschehen. Gewisse Zustände als eher zutreffend ansehen bedeutet, daß nach sorgfältigem Abwägen die in Frage kommenden Zustände mit subjektiven Wahrscheinlichkeiten belegt werden. Diese subjektiven Wahrscheinlichkeiten sollen im folgenden als *Plausibilitäten* bezeichnet werden. Aus diesen Plausibilitäten resultieren i.allg. Handlungen, aus diesen gewisse Ergebnisse. Die Verteilung von Plausibilitäten ist in diesem Zusammenhang *kohärent*, wenn die aus ihnen resultierenden Handlungen nicht zwangsläufig zu unliebsamen Konsequenzen führen. Unliebsame Konsequenzen sind solche, die das so handelnde Subjekt zu vermeiden trachtet. Sind alle Konsequenzen unliebsam, so versucht S wenigstens, den Schaden möglichst gering zu halten. In Kapitel 4 wird untersucht, wie S auf Grundlage dieser Plausibilitäten eine Handlung auswählen kann, deren Konsequenzen für S möglichst annehmbar sind.
Ich beginne mit der Betrachtung von Plausibilitäten von Zufallsgrößen, worin sich die Behandlung von Ereignissen problemlos einbetten läßt, weil Ereignisse als spezielle Zufallsgrößen aufgefaßt werden können (s.o.). Die folgenden Ideen finden sich in klarer Form bereits in Kants "Kritik der praktischen Vernunft", wo es heißt: „Der gewöhnlichste Probierstein: ob etwas bloße Überredung, oder wenigstens subjektive Überzeugung ist, d.i. festes Glauben sei, was jemand behauptet, ist das Wetten. Öfters spricht jemand Sätze mit so zuversichtlichem und unlenkbarem Trotze aus, daß er alle Besorgnis des Irrtums gänzlich abgelegt zu haben scheint. Eine Wette macht ihn stutzig. Bisweilen zeigt sich, daß er zwar Überredung genug, die auf einen Dukaten an Wert geschätzt werden kann, aber nicht auf zehn, besitze. Denn den ersten wagt er noch wohl, aber bei zehnen wird er allererst inne, was er vorher nicht bemerkte, daß es nämlich doch wohl möglich sei, er habe sich geirrt." [Ka, S.B852f]

1.3.1 Einschätzung von Zufallsgrößen

Betrachten wir[2] zunächst den Fall des *zufälligen Gewinns* X, d.h. einer Zufallsgröße, die die Bedeutung eines (evtl. negativen) Gewinnes für S hat. Der Wert von X ist eine bestimmte reelle Zahl, die S nicht bekannt ist. Dieses Subjekt S fragen wir, welchen sicheren Gewinn es als gleichwertig zu X betrachtet. Diesen Wert nennen wir den *Preis von* X und bezeichnen ihn mit $p(X)$. Unter normalen Umständen ist der Preis nicht additiv, d.h. wenn S bereit ist, für X $p(X)$ und für Y $p(Y)$ zu bezahlen, so nicht unbedingt für $X + Y$ $p(X) + p(Y)$. In der Einführung subjektiver Einschätzung, wie sie weiter unten erfolgt, wird diese Eigenschaft allerdings grundlegend für kohärentes Verhalten sein. Ein Subjekt handelt kohärent, wenn es sich durch seine Handlung nicht einen voraussehbaren Schaden einhandelt. Dieser Schaden könnte zum Beispiel darin bestehen, daß S die Differenz zwischen dem gewetteten und dem tatsächlichen Wert bezahlen muß. In diesem Fall wird S sich um eine möglichst genaue Einschätzung bemühen. Zwei Eigenschaften werden von der Preisfunktion verlangt:

1. Der Preis p ist eine additive Funktion

$$p(X + Y) = p(X) + p(Y)$$

2. Der Preis p erfüllt die Ungleichung

$$\inf X \leq p(X) \leq \sup(X)$$

   Dies ist natürlich nur eine Beschränkung, wenn X beschränkt ist.

3. aus 1. und 2. folgt, daß p eine lineare Funktion ist, d.h.

$$p(aX) = ap(X) \quad \forall a \in \mathbb{R}$$

   und allgemeiner

$$p(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1p(X_1) + a_2p(X_2) + \ldots + a_np(X_n)$$

   für endlich viele Summanden

Um die Preisfunktion genauer zu fassen, werde ich nun zwei Kriterien angeben, sowie die Bedingungen für eine kohärente Wahl, wenn S Verluste durch seine Einschätzung vermeiden will.

Erstes Kriterium: Gegeben sei eine Zufallsgröße X. S muß einen Wert $\bar{x}$ wählen, wonach eine Wette mit dem Gewinn $c(X - \bar{x})$ durchgeführt wird, wobei der Wert c beliebig vom Wettgegner gewählt werden kann.

[2]Um den Leser in die Entwicklung der folgenden Gedanken miteinzubeziehen, verwende ich die “wir-“Form

Definition 1.1 *p(X) als subjektive Einschätzung der Zufallsgröße X ist genau der von S gewählte Wert $\bar{x}$.*

Dies ist ähnlich wie der Vorschlag, daß die erste Person S_1 einen Kuchen teilt, die zweite, S_2, allerdings die erste Wahl treffen darf. S_1 wird den Kuchen möglichst gerecht teilen, weil S_2 sowieso das größere Stück nehmen wird.

Kohärenz: Es wird davon ausgegangen, daß S keine Wette eingehen wird, die mit Sicherheit zum Verlust führt, ein sogenanntes "Dutch Book". Eine Menge von Einschätzungen ist *kohärent*, wenn in allen Wettkombinationen keine vollständig negativ ist. Dies heißt nicht, daß keine Verluste möglich sind, wenn die Wette ungünstig verläuft, aber sie treten nicht zwangsläufig auf, durch willentliche Fehleinschätzung zum Beispiel. [3] Mathematisch heißt dies, daß die Werte $\bar{x}_i = p(X_i)$ so gewählt werden müssen, daß es keine Linearkombination

$$Y = c_1(X_1 - \bar{x}_1) + c_2(X_2 - \bar{x}_2) + \ldots + c_n(X_n - \bar{x}_n)$$

mit $\sup Y < 0$ geben kann.

zweites Kriterium: S wird mit der Sanktion L bestraft, wobei L proportional zu dem Quadrat der Differenz zwischen X und dem geschätzten Wert $\bar{x}$ ist, letzterer ist frei wählbar:

$$L = \left(\frac{X - \bar{x}}{k}\right)^2$$

Der Nenner k ist willkürlich und kann von Fall zu Fall variieren. Er dient im wesentlichen dazu, daß L eine Zahl wird, falls die Einheit von X eine andere Dimension hat.

Definition 1.2 *p(X), die subjektive Einschätzung von X, ist der Wert $\bar{x}$, der zu diesem Zweck gewählt wird, sozusagen das kleinste Übel.*

Kohärenz: Es wird davon ausgegangen, daß S keine Sanktion wählen wird, wenn es eine andere gibt, die mit Sicherheit kleiner ist. Eine Menge von Einschätzungen ist kohärent, wenn es keine möglich Wahl gibt, die die Summe der Sanktionen vermindern würde. Das heißt, daß es keine Werte x_i^* gibt, für die die Sanktion

$$L^* = \sum_i \left(\frac{X_i - x_i^*}{k_i}\right)^2$$

immer kleiner ist als

$$L = \sum_i \left(\frac{X_i - \bar{x}_i}{k_i}\right)^2,$$

[3] Ich weise an dieser Stelle darauf hin, daß hier nicht versucht werden soll, zu formalisieren, wie S auf $\bar{x}$ kommt, ob durch simples Raten, durch Auswerten von Zahlentabellen, o.ä. Wichtig ist, daß S möglichst genau seine Einschätzung über X angibt.

für alle möglichen $(X_1, X_2, \ldots, X_n) \in Q$.

Satz 1.1 *Beide Kriterien sind äquivalent, führen also zu der selben Einschätzung*

□

Beweis: Sei $\overline{x}$ die Einschätzung gemäß des ersten Kriteriums, $\overline{\overline{x}}$ die des zweiten.

1. Im ersten Fall wird der zufällige Gewinn X als gleich angesehen wie $\overline{x}$, also jedem $x < \overline{x}$ vorgezogen, aber keinem $x > \overline{x}$.
2. Im zweiten Fall wird der negative Gewinn $-(X-\overline{\overline{x}})^2$ jedem anderen $-(X-x)^2$ vorgezogen, wenn $x \neq \overline{\overline{x}}$. Der Gewinn

   $$G = (X-x)^2 - (X-\overline{\overline{x}})^2$$

   wird also der 0 vorgezogen.

Allgemeiner lassen sich die Sanktionen vergleichen zwischen verschiedenen Werten von x, z.B. $x = a$ und $x = b$. Mit $c = \frac{1}{2}(a+b)$ bezeichne ich den Mittelpunkt des Intervalles $[a, b]$. Die Wahl a wird b vorgezogen, wenn der Gewinn

$$G = (X-b)^2 - (X-a)^2$$

der 0 vorgezogen wird. Das heißt:

$$G = (X^2 - 2bX + b^2) - (X^2 - 2aX + a^2) = 2(a-b)X - (a^2 - b^2) = 2(a-b)(X-c)$$

wird dem 0-Gewinn vorgezogen. Daher ist $p(G) > 0$. Mit dem ersten Kriterium heißt dies, daß $p(G) = 2(a-b)(\overline{x}-c)$ ist. Dieser Ausdruck ist positiv, wenn $a > b$ und $\overline{x} > c$ ist, oder wenn $a < b$ und $\overline{x} < c$ gilt. In jedem Fall liegt $\overline{x}$ im Teilintervall $[c, a]$, d.h. näher an a als an b. Als direkte Folgerung des Gesagten ergibt sich als optimale Wahl für x der Wert $\overline{\overline{x}} = \overline{x}$. Die Äquivalenz der Kohärenzbedingungen ergibt sich ebenso.

■

Eine geometrische Interpretation kohärenter Einschätzungen, die hier allerdings nicht hergeleitet werden soll, ist die, daß die Einschätzung von n Zufallsgrößen $X_1, X_2, \ldots, X_n$ gerade die abgeschlossene konvexe Hülle der Menge Q möglicher Werte von $X_1, X_2, \ldots, X_n$ ist. Die subjektive Einschätzung ist so etwas wie der Erwartungswert der Zufallsgröße. Da Ereignisse Zufallsgrößen sind, die nur die beiden Werte 0 und 1 annehmen können, ist die Einschätzung von Ereignissen ein Spezialfall des bisher behandelten. Die subjektive Einschätzung von Ereignissen bezeichne ich im weiteren als *Plausibilität*.

1.3.2 Plausibilitäten

Sind zwei Ereignisse A und B unvereinbar, dann stimmen ihre arithmetische und ihre logische Summe überein. $E = A + B = A \vee B$, so daß bei additivem p gilt: $p(E) = p(A) + p(B)$. Dies gilt natürlich auch für die arithmetische und logische Summe jeder endlichen Anzahl unvereinbarer Ereignisse: $E = E_1 \vee E_2 \vee \ldots \vee E_n = E_1 + E_2 + \ldots E_n$ und somit gilt

Satz 1.2 *Im Falle n unvereinbarer Ereignisse $E_1, E_2, \ldots, E_n$ gilt:*

$$p(E) = p(E_1) + p(E_2) + \ldots + p(E_n)$$

□

als direktes Korollar gilt

Satz 1.3 *In einer endlichen Einteilung ist diese Summe 1*

□

Insbesondere ist bei zwei komplementären Ereignissen E und $\neg E$ diese Summe 1: $p(E) + p(\neg E) = 1$, somit ist $p(\neg E) = 1 - p(E) = \neg p(E)$. Daher

Satz 1.4 *Die Plausibilität zweier komplementärer Ereignisse ist selber komplementär.*

□

Für beliebige Ereignisse muß folgendes gelten:

$$E = E_1 \vee E_2 \vee \ldots \vee E_n = 1 \wedge (E_1 + E_2 + \ldots + E_n) \leq E_1 + E_2 + \ldots + E_n$$

und somit

$$p(E) \leq p(E_1) + p(E_2) + \ldots + p(E_n) = p(E_1 + E_2 + \ldots + E_n).$$

Das eigentliche Problem ist die Wahl kohärenter Plausibilitäten. Hier gilt der folgende

Satz 1.5 (Hauptsatz der Plausibilitäten) *Gegeben seien die Plausibilitäten $p_i = p(E_i)$ $(i = 1, 2, \ldots, n)$ einer endlichen Anzahl von Ereignissen. Die Plausibilität $p(E)$ eines zusätzlichen Ereignisse ist entweder*

(i) bestimmt durch die p_i, wenn E von den E_i linear abhängig ist, oder

(ii) kann an diese kohärent angefügt werden.

□

Der Beweis ist nicht schwierig, bedarf aber einiger begrifflicher Vorarbeit, die für das weitere nicht wesentlich ist. Hierfür verweise ich auf das Buch von De-Finetti, [Df, S.112]. Das Interessante an diesem Satz ist, daß die Kohärenz auch dann nicht verlorengeht, wenn abzählbar unendlich viele Ereignisse betrachtet werden. Allerdings dauert das Anfügen hier unendlich lange. Als notwendige Bedingung für Kohärenz ist hierbei die Additivität der Plausibilitätsfunktion gegeben, womit ihre Forderung auch formal gerechtfertigt ist.

1.4 Bedingte Plausibilitäten

Eine weiterer wichtiger Begriff ist die Einführung von bedingten Plausibilitäten, d.h. die Plausibilität für das Ereignis E, wenn Erkenntnis über das Ereignis H vorhanden ist. Bisher hieß $p(E)$ immer $p(E|H_0)$, wobei H_0 alle Informationen bedeuten, die S zu Zeitpunkt der Einschätzung hatte. Dies ist allerdings so ungenau, daß darauf verzichtet wird, es in dem Symbol H_0 festzuhalten. $p(E|H)$ bedeutet also die Plausibilität, die S dem Ereignis E einräumt, wenn S erfährt, daß H eingetreten ist. Interpretiert auf die beiden Kriterien bedeutet dies: (1) Indem der Preis $p(HE)$ bezahlt wird, kann S sicher sein, eine Zahlungseinheit zu bekommen, wenn HE eintritt. Das gleiche wäre aber auch erreichbar, indem $p(E|H)$ nur dann gezahlt wird, wenn H wahr ist. Also gilt

$$p(HE) = p(H)p(E|H)$$

Dies gilt natürlich auch für eine Zufallsgröße X, anstelle des Ereignisses E. Aus dem zweiten Kriterium ergibt sich folgende

Definition 1.3 *Gegeben sei eine Zufallsgröße X und ein mögliches Ereignis H. Wenn S ein $\overline{x}$ so wählen kann, daß es eine Sanktion der Größe*

$$L = H\left(\frac{X - \overline{x}}{k}\right)^2$$

erleidet, so wird das von S gewählte $p(X|H) = \overline{x}$ die bedingte Plausibilität von X unter der Bedingung H *genannt. Die Sanktion L ist 0, wenn H nicht eintritt.*

Kohärenz bedeutet in diesem Fall, daß S keine Sanktion wählt, wenn es eine Möglichkeit gibt, diese mit Sicherheit zu verringern.

Satz 1.6 *Eine notwendige und hinreichende Bedingung für Kohärenz in der Bestimmung von $p(X|H)$, $p(H)$ und $p(HX)$ ist das Erfülltsein der Gleichung*

$$p(HX) = p(H) \cdot p(X|H)$$

zuzüglich natürlich $\inf(X|H) \leq p(X|H) \leq \sup(X|H)$*, sowie* $0 \leq p(H) \leq 1$*. Ist X ein Ereignis E, so gilt die Beziehung*

$$p(HE) = p(H) \cdot p(E|H).$$

Diese Beziehung heißt dann auch **Satz der bedingten Plausibilität**. *Die entsprechenden Ungleichungen lauten hierbei* $0 \leq p(E|H) \leq 1$ *und* $0 \leq p(H) \leq 1$.

□

Beweis: Zunächst betrachte ich den Spezialfall von Ereignissen, und bezeichne mit x, y, z die nach einem Kriterium gewählten Plausibilitäten von $p(E|H), p(H)$ und $p(HE)$. Der Satz formuliert sich dann in der Form $z = xy$. Die drei Plausibilitäten x, y, z werden nach dem zweiten Kriteriem sanktioniert, wobei S die Summe dieser Sanktionen möglichst klein halten will. Die Frage ist, wie die Beziehung zwischen diesen drei Größen sein muß, damit die Sanktion minimal ist. Es wird sich zeigen, daß diese Beziehung gerade von der Form $z = xy$ sein wird. Die Sanktion ist mit $k = 1$

$$L = H \cdot (E - x)^2 + (H - y)^2 + (HE - z)^2,$$

hier müssen drei Fälle unterschieden werden:

(i) HE $(H = E = HE = 1)$

(ii) $H\neg E$ $(H = 1, E = HE = 0)$

(iii) $\neg H$ $(H = HE = 0)$

und es ergibt sich:

$$\begin{array}{llllllll} (i) & HE & : & L = u = & (1-x)^2 & + & (1-y)^2 & + & (1-z)^2 \\ (ii) & H\neg E & : & L = v = & x^2 & + & (1-y)^2 & + & z^2 \\ (iii) & \neg H & : & L = w = & & & y^2 & + & z^2 \end{array}$$

Faßt man die Werte x, y, z als Cartesische Koordinaten auf, so sind die Sanktionen u, v, w in den drei Fällen das Quadrat der Abstände von dem Punkt $(1, 1, 1)$ im ersten Fall, von $(0, 1, 0)$ im zweiten, sowie von der x-Achse im dritten Fall, also vom Punkt $(x, 0, 0)$. Die vier Punkte $(x, y, z), (1, 1, 1), (0, 1, 0)$ und $(x, 0, 0)$ liegen dann in einer Ebene, wenn ein fünfter, nämlich $(x, 1, z/y)$, auch auf ihr liegt. Dieser Punkt ist der Schnittpunkt der Geraden durch (x, y, z) und $(x, 0, 0)$ mit der Ebene $y = 1$. Damit muß der Punkt $(x, 1, z/y)$ identisch sein mit dem Punkt $(x, 1, x)$, der auf der Geraden durch $(1, 1, 1)$ und $(0, 1, 0)$ liegt. Daher muß $z = xy$ gelten und der Punkt (x, y, z) muß auf diesem Paraboloiden und natürlich innerhalb des Einheitswürfels liegen. In diesem Fall ist es nicht möglich, die drei Abstände gleichzeitig zu verringern, was in anderen Fällen möglich wäre. Im allgemeineren Fall einer Zufallsgröse X bezeichnen x, y, z die drei Plausibilitäten $p(X|H), p(H)$ und $p(HX)$. Die oben angeführten Argumente sind zwar weiterhin gültig, außer den beiden Punkten $(1, 1, 1)$ und $(0, 1, 0)$ auf der Geraden $y = 1, z = x$ müssen nun alle Punkte in Betracht gezogen werden, deren x-Werte mögliche Werte für die Zufallsgröße X sind und die in Übereinstimmung mit H stehen. Dies ergibt:

$$\begin{aligned} L &= H(X - x)^2 + (H - y)^2 + (HX - z)^2 \\ &= H[(X - x)^2 + (1 - y)^2 + (X - z)^2] + (1 - H)(y^2 + z^2). \end{aligned}$$

Liegt (x, y, z) nicht auf dem Paraboloiden $z = xy$ und somit auch nicht auf der Ebene durch die Gerade $y = 1, z = x$ und den Punkt $(x, 0, 0)$, so liesse sich gleichzeitig der Abstand zur x-Achse als auch zu der Geraden verkleinern. Da dies keine kohärente Wahl gewesen wäre, muß er sowohl auf dem Paraboloiden $z = xy$, als auch innerhalb der Grenzen

$$0 \leq y \leq 1,\ \inf(X|H) \leq x \leq \sup(X|H)$$

liegen.

■

Wegen der Kommutativität des logischen Produktes ist

$$p(HE) = p(EH) = p(E)p(H|E) = p(H)p(E|H)$$

Eine direkte Folgerung ist der

Satz 1.7 (Satz von Bayes)

$$p(E|H) = p(E)\frac{p(H|E)}{p(H)},\ \textit{falls}\ p(H) \neq 0$$

□

Dieser Satz ist auch bekannt unter dem Namen „Schluß von der Wirkung auf die Ursache“, weil die Plausibilität $p(E|H)$ angibt, wie stark S davon überzeugt ist, daß das Ereignis E für das Eintreffen von H verantwortlich war, wenn H tatsächlich eingetroffen ist.

1.5 Stochastische Unabhängigkeit

Mit Hilfe des Begriffes der bedingten Plausibilität ist es auch möglich, den Begriff der stochastischen Unabhängigkeit einzuführen: Die Plausibilität $p(E|H)$ ist entweder gleich $p(E)$, sie ist größer oder kleiner. Dies bedeutet, daß, wenn H eingetreten ist, S entweder seine Meinung über E nicht ändert, E als weniger oder als mehr plausibel annimmt. Im ersten Fall heißt E *stochastisch unabhängig von* H, im zweiten *stochastisch abhängig von* H. Diese Eigenschaft ist symmetrisch, d.h.

$$\frac{p(E|H)}{p(E)} = \frac{p(H|E)}{p(H)} = \frac{p(EH)}{p(E) \cdot p(H)}$$

Es läßt sich also feststellen, daß das Verhältnis, in dem die Plausibilität von E zu- oder abnimmt, wenn E von H abhängt, das gleiche ist, wie das Verhältnis, wenn H von E abhängt. Dieses Verhältnis ist gleich dem zwischen der Plausibilität von EH mit der des Produktes der Plausibilitäten von E und H. Im Falle stochastischer Unabhängigkeit ist dieses Produkt gleich $p(EH)$. Also gilt

$$p(EH) = p(H) \cdot p(E|H) = p(H)p(E),\ \text{falls}\ p(E|H) = p(E)$$

Wenn E und H abhängig oder unabhängig voneinander sind, so gilt dies natürlich auch für $\neg E$ und H, für E und $\neg H$ und für $\neg E$ und $\neg H$. Die Ereignisse $E_1, E_2, \ldots, E_n$ sind dann *stochastisch unabhängig*, wenn

$$p(E_{i_1} E_{i_2} \cdots E_{i_k}) = p(E_{i_1}) p(E_{i_2}) \cdots p(E_{i_k})$$

gilt, für beliebige Produkte von $k \leq n$ Ereignissen $E_{i_1}, E_{i_2} \cdots, E_{i_k}$. Auch hier können die E_i beliebig durch ihr Komplement $\neg E_i$ ersetzt werden.

1.6 Verteilungen

Es ist an dieser Stelle gar nicht notwendig, das Konzept der Verteilung vollständig einzuführen. Ich beschränke mich also nur auf die Verteilungen, die im folgenden noch gebraucht werden, zum einen die endlichen Verteilungen auf einer Einteilung und die Verteilungen auf einem beschränkten reellen Intervall.

1.6.1 Endliche Verteilungen

Gegeben sei eine endliche Einteilung A der möglichen Weltzustände, d.h. es existiert eine endliche Anzahl an Ereignissen $A = \{a_1, a_2, \ldots, a_n\}$ von denen genau eines zutrifft. Eine *Plausibilitätsverteilung* v auf A ist eine Funktion $v : A \to [0, 1]$, die jedem Ereignis aus A eine Plausibilität zuordnet, derart, daß die Summe all dieser Plausibilitäten 1 ergibt:

$$\sum_{i=1}^{n} v(a_i) = 1$$

Die Funktion v muß hierbei additiv sein, d.h.

$$v(a_1 + a_2) = v(a_1) + v(a_2)$$

1.6.2 Verteilungen auf [a,b]

Gegeben sei ein reelles Intervall $I = [a, b],\ -\infty < a \leq b < \infty$ und eine rechtsseitig stetige, monoton steigende Funktion $F : \mathbb{R} \to [\not\vdash, \not\dashv]$ mit $F(a) = 0,\ F(b) = 1$. Eine solche Funktion heißt *Verteilungsfunktion auf* $[a, b]$. Eine Interpretation ist die der Massenverteilung auf $[a, b]$. Für $x \in [a, b]$ gibt $F(x)$ die Masse links von der Stelle x an , $1 - F(x)$ die Masse rechts davon. Ist eine Masse p_h am Punkt x_h konzentriert, so hat F an x_h eine Sprungstelle der Höhe $F(x_h + 0) - F(x_h - 0)$. Eine Verteilung mit nur konzentrierten Massen und $\sum_h p_h = 1$ heißt *diskrete Verteilung*. Eine Ausartung ist die Einpunktverteilung, die die gesamte Masse in einem Punkt x_0 vereinigt. Eine Verteilung ohne Massenkonzentration heißt *stetige Verteilung*. Ist F rechtsseitig differenzierbar, so gibt es eine Funktion f, die *Dichtefunktion von* F, so daß die Masse sich bestimmt aus

$$F(x) = \int_a^x f(x) dx.$$

Natürlich kann die Gesamtmasse den Wert 1 nicht übersteigen:

$$\int_a^b f(x)dx = 1.$$

Der mathematische Erwartungswert einer Verteilung F auf $[a, b]$ ergibt sich dann durch

$$E(F) = \int_a^b xf(x)dx.$$

Für eine diskrete Verteilung mit den Sprungstellen $x_1, x_2, \ldots, x_n$ und den zugeordneten Massen $p_1, p_2, \ldots, p_n$ ergibt sich der Erwartungswert entsprechend aus

$$E(F) = \sum_{i=1}^{n} x_i p_i.$$

Der subjektive Erwartungswert einer Zufallsvariablen ist weiter oben bereits eingeführt worden als Einschätzung unter Androhung von Verlust oder Sanktionen.

1.7 Axiomatik

Die gesamte Einführung in die Wahrscheinlichkeitsrechnung aus der Sichtweise subjektiver Einschätzung basiert im wesentlichen auf drei Eigenschaften, die im folgenden als Axiome eingeführt werden. Wird davon ausgegangen, daß ein Subjekt sich nicht absichtlich selber schädigen will, so können diese Axiome als Grundlage für rationale Einschätzung angesehen werden, weil nur bei ihrer Einhaltung kohärentes Verhalten möglich ist.

1. **nicht-Negativität: Gilt sicher $X \geq 0$, so ist sicher $p(X) \geq 0$**

2. **endliche Additivität der Plausibilitätsfunktion p:**

$$p(X + Y) = p(X) + p(Y)$$

Daraus folgt sofort

$$p(aX) = ap(X), \;\; \inf X \leq p(X) \leq \sup(X), \;\; a \in \mathbb{R}$$

sowie die Konvexitätsbedingung, die die Axiome 1 und 2 enthält:

Jede lineare Gleichung oder Ungleichung zwischen Zufallsgrößen X_i muß durch die Plausibilitäten $p(X_i)$ erhalten werden:

gilt sicher $c_1X_1 + c_2X_2 + \ldots + c_nX_n = c($ oder $\geq c)$
so folgt notwendig $c_1p(X_1) + c_2p(X_2) + \ldots + c_np(X_n) = c$ (oder $\geq c)$

Anders ausgedrückt: Keine Linearkombination von Zufallsgrößen kann immer

positiv sein, d.h. die $p(X_h)$ müssen so gewählt werden, daß für alle $c_1, c_2, \ldots, c_n$ kein $c > 0$ existiert, derart, daß

$$c_1(X_1 - p(X_1)) + c_2(X_2 - p(X_2)) + \ldots + c_n(X_n - p(X_n)) \geq c$$

mit Sicherheit erfüllt ist.
Dies ist die Bedingung für Kohärenz; eine Einschätzung p, die diese Axiome erfüllt, heißt kohärent.

Das dritte Axiom bezieht sich auf bedingte Plausibilitäten:

3. Für bedingte Plausibilitäten gilt folgende Ungleichung:

$$p(E|H) = p(EH)/p(H), \;\; p(X|H) = p(XH)/p(H)$$

Dieses dritte Axiom garantiert die Kohärenz von bedingten Einschätzungen, so daß alle drei Axiome als Ausdruck rationalen Verhaltens aufgefaßt werden können. In den folgenden Kapiteln werden noch andere Axiome hinzugefügt werden, deren Einhaltung als Bedingung für Rationalität angesehen werden kann.

2 Sachinformationsfragen und Information

2.1 Heuristik

Ausgehend von einer mehr intuitiven Beschreibung, was eine Frage sein kann, werde ich im Laufe des Kapitels den Komplex der Fragen auf Sachinformationsfragen einschränken. Das sind Fragen, die nur eine endliche Zahl verschiedener Antworten zulassen, die dem fragestellenden Subjekt S alle bekannt sind. Die Sachinformationsfrage versucht lediglich zu ermitteln, welche Antwort tatsächlich zutrifft. Dieser Fragetypus erlaubt eine mathematisch geschlossene Darstellung, die darüber hinaus den Vorteil hat, formal eng mit der mathematischen Beschreibung von Entscheidungssituationen verwandt zu sein.
Unter Information verstehe ich alles, was Antwort auf eine Sachinformationsfrage sein kann. Dies ist keine intuitive Begriffsbestimmung, sondern tatsächlich eine operationale Definition, die ich kurz anderen Informationsbegriffen gegenüberstellen will.
Die Menge aller Sachinformationsfragen sei $\mathbb{F}$, die Menge aller Sätze, die Antworten auf eine Sachinformationsfrage sein können, bezeichne ich mit $\mathbb{A}$. Ein Satz einer natürlichen Sprache trägt nur dann Informationen für ein Subjekt S, wenn dieses eine Frage hatte, die mit diesem Satz beantwortet wird. Zwar kann so praktisch jeder Satz Informationen tragen, aber nicht alle sind Antworten auf Fragen von S. Somit ist der Begriff der Information abhängig vom Vorwissen des jeweiligen Subjekts und liegt den Dingen nicht bei. Dies ist vergleichbar mit der Verneinung der objektiven Wahrscheinlichkeit im letzten Kapitel. Der Shannonsche Informationsbegriff basiert auf dieser objektiven Wahrscheinlichkeit, somit wird Information ebenfalls zu einer Eigenschaft der Welt.

Information wird neben Energie und Materie als eine der Urgrößen des Universums angesehen. Meines Erachtens beruht diese Annahme auf der Möglichkeit der Anwendung informationstheoretischer Methoden, insbesondere des Begriffes der Entropie, auf statistische Beschreibungen natürlicher Vorgänge. Die Unsicherheit, die in diesen Beschreibungen steckt, bedingt nun, daß das Eintreten eines bestimmten Ereignisses um so mehr Information trägt, je unwahrscheinlicher sein Eintreffen war. Im Zusammenhang mit meinem subjektiven Informationsbegriff bedeutet dies, daß das Eintreffen eines Ereignisses für das Subjekt S unwahrscheinlich war. Erst wenn S wissen wollte, ob besagtes Ereignis zutrifft, trägt sein Eintreten Information. Wird von diesem Subjekt abstrahiert, dann läßt sich durchaus Information als Grundgröße dieses Ereignisses interpretieren. Dies impliziert allerdings die Annahme, daß das Universum seinem Wesen nach statistisch ist. Die menschliche Unfähigkeit, natürliche Vorgänge in ihrer Totalität zu erfassen und vollständig zu beschreiben wird so aus dem Menschen herausprojiziert und zu einer Eigenschaft der den Menschen umgebenden Umwelt gemacht. Wenn ein Gemälde schlecht ist, kann man es aber nicht dem Motiv anlasten, daß der Maler schlecht ist oder schlechte Materialien benutzt. Die statistische Welterklärung ist vielmehr ein Ausdruck menschlicher Erkenntnisfähigkeit oder -unfähigkeit.
Überall wo diskrete statistische Verteilungen benutzt werden, lassen sich die Begriffe der Informationstheorie anwenden. Der Betrachter wird von der so erklärten Natur abstrahiert, Information steckt "in den Sachen selber". Ich empfinde die Bezeichnung 'Information' für diese Abstraktion als ungeschickt[4] und möchte weiter an einem subjektiven Informationsbegriff festhalten. Dieses Problem ist eine moderne Fassung des Rationalismus/Empirismus-Streits und kann prinzipiell auch ähnlich aufgehoben[5] werden, wie Kant es in seiner 'Kritik der praktischen Vernunft' durchgeführt hat. Der Mensch ist nicht Beobachter der Natur, sondern schreibt ihr durch seinen Erkenntnisapperat die Struktur vor. Nicht die Erkenntnis richtet sich nach den Gegenständen, sondern vielmehr die Gegenstände nach der Erkenntnis, vgl.[Ka, S.BXVI]. Somit ist Information als Ausdruck der statistischen Beschaffenheit, ebenso wie Raum und Zeit zwar eine Größe der Natur, nicht aber eines 'Dinges an sich'. Jede Erkenntnis fängt im Menschen an, der die Natur vor jeder Erkenntnis, also a priori im eigentlichen Sinne, vorstrukturiert. Information bleibt gekoppelt am erkennenden Subjekt und wird gleichzeitig Grundgröße der Natur. Deswegen betrachte ich im weiteren Verlauf die Information aus der subjektivistischen Perspektive, operational eingeführt als Antworten auf bestimmte Fragen eben dieses erkennenden Subjekts. Dabei gehe ich davon aus, daß dieses Subjekt grundsätzlich in der Lage ist, seinen Informationsbedarf in Worte zu fassen. Die erhaltene Antwort beseitigt dann eine Unsicherheit, Information ist beseitigte Unsicherheit.
Um den Versuch zu umgehen, menschliche Sprache zu formalisieren, gehe ich

[4]Shannon hat selber für seine Theorie die Bezeichnung "Signaltheorie" vorgeschlagen.

[5]ganz im Sinne einer Hegelschen Dialektik, aufheben im dreifachen Sinne von 'auflösen', 'hochheben' und 'bewahren'.

im folgenden von einem sprechfähigen Subjekt S aus, das in der Lage ist, zu entscheiden, ob zwei Sätze der Sprache für S selbst das gleiche bedeuten. Mathematisch beschreiben läßt sich dies durch die Äquivalenzrelation "ist semantisch gleich wie", im folgenden abgekürzt mit "$\sim$", fassen. So bedeutet z.B. der Satz „Am Samstag tragen alle Schornsteinfeger lila Krawatten" für bestimmte Sprachgemeinschaften das gleiche wie „Sonnabends werden von allen Rauchfangfegern violette Schlipse umgebunden".
Eine Frage läßt sich nun beschreiben als eine Einteilung von $\mathbb{A}$ in Äquivalenzklassen, geschrieben als $\mathbb{A}/\mho$. Zu jeder Frage kann S entscheiden, ob zwei Antworten a_1 und a_2 aus $\mathbb{A}$ das gleiche aussagen. Auf die Frage "Welche Farbe hatte das Auto ?" gilt

„Das Auto war grün" $\sim$ „grün"

Beide Aussagen können in anderen Kontexten, d.h. Fragesituationen zu ganz anderen Aussagen äquivalent sein " z.B. auf die Frage "Welches ist Deine Lieblingsfarbe ?", wo die Antwort "Das Auto war grün" ohne Sinn ist, die Antwort "grün" aber äquivalent zu "Meine Lieblingsfarbe ist grün" ist. Für S können aber auch Aussagen wie "es war tannengrün" und "es war flaschengrün" äquivalent sein zu "es war grün". In diesem Fall erhielt S mehr Informationen als es verarbeiten konnte und wird auch nur behalten: "Das Auto war grün". Die Farbpalette von S ist noch nicht hinreichend differenziert, was durch geeignete Schulung sicherlich zu verbessern wäre. Insofern spiegeln Fragen immer den gegenwärtigen Informationsstand des fragenden Subjekts wider, indem gewisse Antworten als gleich angesehen werden.
Dieser Informationsstand, aber auch die Fragen, die tatsächlich als Problem empfunden werden, ist in hohem Maße vermittelt durch die Einbindung des Subjekts in eine Kommunikationsgemeinschaft. Insofern spiegeln Fragen auch das Selbstverständnis dieser Gemeinschaft wider. Die sogenanten 'großen Wissenschaftler' haben es immer wieder geschafft, ihre Fragen zu Fragen der sie umgebenden Gemeinschaft zu machen (und sie oft auch zu beantworten versucht). Ob die Erde sich um die Sonne dreht, war seit den Griechen keine Frage, bis Keppler nicht nur sein Leben riskierte, um diese Frage zu stellen, sondern sie auch in einer Weise beantwortete, die das herrschende Weltbild umstürzte. Dieser Aspekt des Problemdrucks einer Frage wird im vierten Kapitel wieder aufgegriffen, wenn es darum geht, den psychologischen Nutzen einer Frage zu beschreiben. Zunächst werde ich die Heuristik noch etwas verfeinern.
Je zwei Fragen f_1, f_2 unterscheiden sich dadurch, daß es in ihnen mindestens zwei Antworten gibt, die in f_1 zueinander äquivalent sind, in f_2 aber nicht. Diese Unterscheidung wird immer gefällt von dem Subjekt S_1 und kann bei einem anderen Subjekt S_2 unterschiedlich ausfallen, z.B. weil S_2 ein differenzierteres Vokabular hat und noch zwischen "Minzgrün", "Tannengrün" und "Flaschengrün" unterscheiden kann. Das Feststellen semantischer Äquivalenz ist eine weitgehend subjektive, allerdings spielen Erziehung und Ausbildung innerhalb der Kommunikationsgemeinschaft eine entscheidende Rolle. Wie S diese Entscheidung fällt,

entzieht sich bislang jeglicher Formalisierung[6] und diese Arbeit will auch keinen Teil dazu beitragen.
Unter einer Sachinformationsfrage verstehe ich eine Frage, auf die der Fragesteller S im voraus eine endliche Anzahl von Antworten kennt und lediglich wissen möchte, welche von diesen Antworten zutrifft. Beispiele hierfür sind Fragen wie „An welchem Terminal fliegt das Flugzeug nach Ottawa ?“ oder „möchten sie etwas trinken ?“ inklusive „Was möchten sie trinken ?“ aber auch „Wie spät ist es bitte ?“. Letztere Frage ließe eigentlich eine überabzählbare Menge an Antworten zu, tatsächlich begnügt sich der Fragesteller aber meistens mit Antworten im Fünf-Minuten-Abstand, präzisere Antworten werden in die entsprechenden vorweg angenommenen Antwort-Schemata eingegliedert. Die Antwort „Es ist zweiundzwanzig nach drei“ wird dann gleichgesetzt mit der Antwort „Es ist zwanzig nach drei“. Angenommen, es ist wirklich 15:22, dann wird das Mehr an Information in der zweiten Antwort unberücksichtigt gelassen. Wird die Uhrzeit in Minutengenauigkeit erwartet, dann wären auch mehr Antwortmöglichkeiten in Betracht gezogen worden. Das Informationsmaß hängt also mit der Anzahl der von S in Betracht gezogenen Antworten zusammen.
Jede der von S als möglich erachteten Antworten a_i wird mit einer Plausibilität belegt, d.h. S rechnet bis zu einem gewissen Grad damit, daß a_i zutrifft.
Selbstverständlich gibt es zahlreiche Fragen, die nicht von vorneherein in Form einer Multiple-Choice-Frage gestellt werden können: „Wie heissen Sie ?“ oder „Was ist Information“, oder Fragen, deren eigentlicher Sinn nicht das Erlangen von Information ist: „Bist Du traurig ?“. Solche Fragen sind aber nicht Gegenstand dieser Arbeit und entziehen sich im Zweifelsfall sowieso einer sinnvollen Beschreibung durch Mathematik.

Der Leser, der sich bei Einführung von Begriffen wie “Menge aller möglichen Antworten“ oder “semantische Äquivalenz“ unwohl fühlen sollte, sei dahingeend beruhigt, daß diese Begriffe nur der Heuristik dienen und zu einem späteren Zeitpunkt wieder entfernt werden.
Eine Antwort beschreibt einen Weltzustand, ein Ereignis. Bei zwei Antworten a_1 und a_2, die S als plausibel einschätzt, wird davon ausgegangen, daß sie unvereinbar sind, trifft a_1 zu so nicht a_2 und umgekehrt. Diese Menge unvereinbarer Antworten soll auch gleichzeitig eine vollständige Einteilung aller Weltzustände sein, d.h. genau eine Antwort trifft zu. Die Antworten dieser Menge heißen *wesentliche Antworten*. Die wesentlichen Antworten sind nicht zu verwechseln mit Antworten, die tatsächlich gegeben werden können, wenn S die Frage stellt. Wesentliche Anworten werden hier als vollständig unvereinbar angesehen.
Für jede Frage $f \in \mathbb{F}$ bezeichne $\psi(f) : \mathbb{F} \to \wp(\mathbb{A}/\mathfrak{V})$ die Menge der wesentlichen

[6] der größte Versuch dieser Formalisierungen finden sich in den Arbeiten der logischen Empiristen, die alle Aussagen auf bestimmte, empirisch überprüfbare Basissätze zurückführen wollten. Unter dem Einfluß von Frege und dem jungen Wittgenstein versuchte der Wiener Kreis, die Zusammensetzung komplexerer Sätze auf der Basis logischer Verknüpfungen. Näheres dazu findet sich in Kapitel 4 dieser Arbeit.

Antworten auf f. Hierbei bezeichnet $\wp(\mathbb{A}/\mathfrak{V})$ die Potenzmenge von $\mathbb{A}/\mathfrak{V}$. Sind $a, b \in \psi(f)$, so ist weder $a + b$, noch $a \cdot b$ in $\psi(f)$. Auf die Frage „Welche Farbe hat das Auto ?“ ist weder „rot oder grün“ noch „rot und grün“ eine plausibele Antwort im Sinne von tatsächlicher Situationsbeschreibung (wenn S mit einem einfarbigem Auto rechnet). Die erste Antwort liefert Information, jedoch nicht genug, die zweite liefert keine brauchbare Information. Die Antwort „es ist rot oder grün“ kann natürlich von einem Menschen gegeben werden, der versucht, sich an die Farbe zu erinnern (oder der rot/grün blind ist), aber das bedeutet nur, daß dieser Mensch sich nicht genau an die Farbe erinnert (oder sie nicht identifizieren kann), nicht aber, daß das Auto tatsächlich die Farbe „rot oder grün“ hat. Die Funktion $\psi(f)$ liefert nur die Antworten, die tatsächlich der Fall sein können und nicht deren mögliche Verknüpfungen.
Die Menge $\psi(f)$ wird als endlich mit der Mächtigkeit $n = |\psi(f)|$ angenommen, $\psi(f) = \{a_i \in \mathbb{A}/\mathfrak{V}; \wr_{\mathfrak{V}}(\eth_{\beth}) > \nvDash, \beth = \nVdash \ldots \ltimes, \sum \wr_{\mathfrak{V}}(\eth_{\beth}) = \nVdash\}$. Das Plausibilitätsmaß p_f ist auf $\psi(f)$ stets echt positiv. Die Zahl n der wesentlichen Antworten auf f heißt *Grad der Frage* f und wird abgekürzt mit $grad(f)$.

2.2 Information

Wir setzen uns nun die Aufgabe, ein Maß für die Information zu finden, die in einer Antwort stecken kann. Dies ist unerläßlich für eine weiter Behandlung mit Hilfe der Mathematik, die, zumindest als Sprache für Modelle, die Wissenschaft vom Messen, Zählen und Wiegen ist.
Eine Funktion ϕ, die die Information beschreibt, die S beim Erlangen einer bestimmten Antwort aus n (gleichplausiblen) Möglichkeiten erhält, sollte nun die folgenden drei Bedingungen erfüllen:

1. $\phi(1) = 0$, d.h., wenn S sowieso nur mit einer Antwort gerechnet hat, so enthält diese auch keine Information.

2. $1 \leq n_1 \leq n_2 \Rightarrow \phi(n_1) \leq \phi(n_2)$, wenn mehr Antwortmöglichkeiten zur Verfügung stehen, so erhöht sich die Information, die bei der Auswahl einer davon gewonnen wird.

3. $\phi(n_1 n_2) = \phi(n_1) + \phi(n_2)$, wenn die erste Frage n_1 Antworten zuläßt, die zweite n_2, und beide Fragen voneinander unabhängige Sachverhalte erfragen, so hat die zusammengesetzte Frage $n_1 \cdot n_2$ Möglichkeiten, beantwortet zu werden. Die somit gewonnene Information ist aber die Summe der Einzelinformationen, also so, als ob die Fragen hintereinander gestellt worden wären.

Eine Funktion, die diese Bedingungen erfüllt, ist eine Logarithmusfunktion mit beliebiger Basis echt größer 1. Dies verleitete Hartley dazu, das Informationsmaß einer Frage mit n Antworten als

$$\log n$$

festzulegen. Allerdings berücksichtigt diese Festlegung nicht die unterschiedlichen Plausibilitäten, mit denen S diese Antworten gewichtet. Sind diese Plausibilitäten jedoch für alle Antworten gleich, so läßt sich $\log n$ auch wie folgt schreiben:

$$\log n = \sum_{i=1}^{n} \frac{1}{n} \log n = -\sum_{i=1}^{n} \frac{1}{n} \log \left(\frac{1}{n} \right)$$

Der *Informationsbedarf der Frage* f läßt sich nun in Analogie dazu beschreiben als

$$H(f) = -\sum_{i=1}^{n} p_f(a_i) \log p_f(a_i) = H(p_f(a_1), \ldots, p_f(a_n)) \tag{1}$$

Diese Formel wurde von Claude Shannon in [Sh] in Anlehnung an die identische Formel in der Thermodynamik als "Entropie" bezeichnet und ist Grundlage der klassischen Informationstheorie.
Die hier gewählte Bezeichnung "Informationsbedarf" deutet darauf hin, daß S eine Frage hat, die Information einfordert. Wenn diese Frage zur vollen Zufriedenheit von S beantwortet ist, so ist dieser Wert im Mittel identisch mit dem Maß an Information, die in der vollen Beantwortung von f steckt. Insofern ließe sich auch von "Information der Frage f" sprechen, was den tatsächlichen Sachverhalt allerdings verzerrt widergibt. Eine Frage kann keine Information enthalten, außer der, daß der Fragesteller Informationsbedarf hat. Daher werde ich die etwas längere aber exaktere Formulierung "Informationsbedarf" beibehalten.
Die Basis des Logarithmus gibt die Einheit an, in der die Information gemessen wird. Ein Logarithmus zur Basis zwei bedeutet, daß die Informationen in *bit* gemessen wird. *'bit'* bedeutet 'binary digit' und wird im allgemeinen als 0 oder als 1 geschrieben. Eine Frage, die nur zwei Antworten zuläßt, z.B. "ja" oder "nein", erfragt, wenn beide Antworten gleich plausibel sind, ein *bit* Information. Ein Logarithmus zur Basis k erfragt in Verallgemeinerung zu *bit* ein *k-it* Information. Im Falle $k = 10$ heißt die Einheit *dit*. Logarithmen lassen sich mit Hilfe der Formel

$$\log_n(x) = \frac{\log_m(x)}{\log_m(n)}$$

ineinander umrechnen.

2.3 Eigenschaften des Informationsmaßes

Im folgenden bezeichne f immer eine Sachinformationsfrage aus $\mathbb{F}$, p das zu f gehörige Plausibilitätsmaß, $\psi(f) \subset \mathbb{A}/\mathfrak{V}$ die Menge der wesentlichen Antworten mit der Mächtigkeit n.

1. (a) Der Wert von $H(p_1, \ldots, p_n)$ ändert sich bei beliebigen Vertauschungen der Werte $p_1, \ldots, p_n$ nicht.
   Dies folgt sofort aus der Kommutativität der Summenbildung

   (b) $H(p_1, \ldots, p_n)$ ist stetig
   Als Konsequenz der Stetigkeit der Logarithmusfunktion

2. $H(f) = 0 \Leftrightarrow \exists i \in \{1, \ldots, n\} : p(a_i) = 1 \wedge \forall j \neq i : p(a_j) = 0$

   Eine Frage, auf die die Antwort dem Fragesteller von vorneherein bekannt ist, also die Plausibilität 1 hat, erfragt keine Information. Die Menge der wesentlichen Antworten enthält dann eigentlich nur ein einziges Element, dennoch soll zugelassen werden, daß sich durch Beantwortung die Plausibilitäten auf den wesentlichen Antworten verschieben. Wird die Frage vollständig beantwortet, so vereinigt eine Antwort alle Plausibilität auf sich. Eine solche Frage wird als die *leere Frage* bezeichnet und als ausgeartete Frage behandelt.

3. ist $p_i \geq 0$ und $q_i \geq 0$ für $i = 1, \ldots, n$, mit $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$, so ist

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i. \tag{2}$$

   Beweis: Es gilt für jede positive reelle Zahl x:

$$\log x \leq (x-1) \log e. \tag{3}$$

   Beweis:

$$\ln x \leq (x-1) \text{ oder auch } f(x) = x - 1 - \ln x \geq 0, x > 0$$

   Die Funktion f ist stetig differenzierbar in $\mathbb{R}^+$, ihr Extremwert liegt bei

$$f'(x) = 1 - \frac{1}{x} = 0 \Rightarrow x = 1, \text{ mit } f(1) = 0$$

$$f''(x) = \frac{1}{x^2} \geq 0$$

   $f(x)$ ist mithin eine konkave Funktion, $f(1) = 0$ ist ihr Minimum. Also gilt $f(x) \geq 0$ für $x > 0$.

Und weiter

$$\sum_{i=1}^{n} p_i \log q_i - \sum_{i=1}^{n} p_i \log p_i = \sum_{\substack{i=1 \\ p_i \neq 0}}^{n} p_i \log \frac{q_i}{p_i} \leq \sum_{\substack{i=1 \\ p_i \neq 0}}^{n} p_i (\frac{q_i}{p_i} - 1) \log e$$

$$= \sum_{\substack{i=1 \\ p_i \neq 0}}^{n} (q_i - p_i) \log e \leq \log e \sum_{i=1}^{n} (q_i - p_i) = \log e \Big(\underbrace{\sum_{i=1}^{n} q_i}_{=1} - \underbrace{\sum_{i=1}^{n} p_i}_{=1} \Big) = 0$$

wegen $p_i = 0 \Rightarrow p_i \log q_i = p_i \log p_i = 0$ werden nur die Summanden mit $p_i \neq 0$ berücksichtigt. ■

4. Die Information ist maximal, wenn die Plausibilität der Antworten gleich ist, bei n Antworten also $\frac{1}{n}$. Diese Plausibilitätsverteilung ist charakteristisch für das Prinzip des fehlenden Grundes, also wenn kein Grund zu der Annahme besteht, daß eine Antwort plausibler ist als eine andere:

$$H(p_1, \ldots, p_n) \leq H(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n-mal}). \tag{4}$$

Beweis: setzt man in (2) $q_i = \frac{1}{n}$, so ergibt sich

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log \frac{1}{n} = -\log \frac{1}{n} \sum_{i=1}^{n} p_i = \log n.$$

Aus der Monotonie der Logarithmus-Funktion folgt, daß bei gleicher Plausibilität der Informationsbedarf mit wachsender Anzahl möglicher Antworten ebenfalls monoton wächst. Dies deckt sich gut mit der Interpretation des Informationsbegriffes durch Ursul, der Information als Vielfalt einführt [Ur]. Zusammen mit Eigenschaft (1) bedeutet Eigenschaft (3), daß erst da Informationen vorliegen, wo mehr als eine Antwort möglich ist, und daß mit wachsender Vielfalt möglicher Antworten auch die Menge an Information zunimmt.

5. Es gilt:

$$\begin{aligned} H(p_1, \ldots, p_n) &= H(p_1, \ldots, p_{i-1}, p_i + p_{i+1}, p_{i+2}, \ldots, p_n) \\ &+ (p_i + p_{i+1}) H\left(\frac{p_i}{p_i + p_{i+1}}, \frac{p_{i+1}}{p_i + p_{i+1}} \right). \end{aligned}$$

Beweis: o.B.d.A betrachte ich den Fall $i = 1$ (wegen Bedingung 1a):

$$H(p_1 + p_2, p_3, \ldots, p_n) = -(p_1 + p_2)\log(p_1 + p_2) - \sum_{i=3}^{n} p_i \log p_i$$

$$\begin{aligned}(p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) &= -p_1 \log \frac{p_1}{p_1 + p_2} - p_2 \log \frac{p_2}{p_1 + p_2} \\ &= (p_1 + p_2)\log(p_1 + p_2) - p_1 \log p_1 \\ &\quad -p_2 \log p_2.\end{aligned}$$

Eine Addition der beiden Gleichungen liefert die Behauptung. Für eine Interpretation muß etwas weiter ausgeholt werden:

Nehmen wir dazu an, die Frage f ließe drei Antworten zu, a_1, a_2, a_3, jeweils mit den Plausibilitäten p_1, p_2, p_3. Der Informationsbedarf dieser Frage ist $H(f) = H(p_1, p_2, p_3)$. Diese Frage ließe sich z.B. stellen in der Form „Welcher der Fälle 'a_1, a_2, a_3' trifft zu ?“ [7] Um die gewünschte Information zu erhalten, kann die Frage auch aufgeteilt werden in zwei Fragen f_1, f_2, von denen f_1 die ersten beiden Antworten zusammenfaßt. Die Frage f_1 lautet dann „trifft '$b = a_1 + a_2$' oder 'a_3' zu ?“, wobei b die Plausibilität $p_1 + p_2$ hat. Der Informationsbedarf dieser Frage ist $H(f_1) = H(p_1 + p_2, p_3)$, was sicherlich kleiner ist als der der Frage f, weil eine Beantwortung nicht in jedem Fall die gleiche Information liefert wie eine Beantwortung von f. Sollte nämlich die Antwort „'b' trifft zu“ gegeben werden, so muß mit Hilfe der Frage f_2 noch ermittelt werden, ob a_1 oder a_2 zutrifft. Der Informationsbedarf dieser Frage ist dann

$$H(f_2) = H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

Der kohärente Fragesteller schätzt mit der Plausibilität $p_1 + p_2$, daß er nach der Frage f_1 auch noch die Frage f_2 stellen muß, um an die gewünschten Informationen zu gelangen. Daher läßt sich annehmen, daß der Informationsbedarf der Frage f um $(p_1 + p_2) \cdot H(f_2)$ größer ist als der von f_1, also

$$H(p_1, p_2, p_3) = H(p_1 + p_2, p_3) + (p_1 + p_2) \cdot H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$

Die selbe Überlegung führt, angewandt auf eine Frage mit n Antworten und der Eigenschft 1(a) zur Eigenschaft 5.

[7] tatsächlich kann jede Sachinformationsfrage in dieser Form gestellt werden.

6.
$$H\Big(\underbrace{\frac{1}{nl},\frac{1}{nl},\ldots,\frac{1}{nl}}_{nl-mal}\Big) = H\Big(\underbrace{\frac{1}{n},\ldots,\frac{1}{n}}_{n-mal}\Big) + H\Big(\underbrace{\frac{1}{l},\ldots,\frac{1}{l}}_{l-mal}\Big) \qquad (5)$$

Beweis: Dies folgt aus

$$H\Big(\underbrace{\frac{1}{nl},\frac{1}{nl},\ldots,\frac{1}{nl}}_{nl-mal}\Big) = log(nl) = log(n) + log(l)$$

Werden zwei Fragen gestellt, die voneinander unabhängige Information erfragen, z.B. „Wie alt sind Sie ?“ und „Welche Farbe haben Ihre Socken ?“, so erfragt die zusammengesetzte Frage „Wie alt sind Sie und welche Farbe haben Ihre Socken“ genauso viel Information, wie die beiden einzelnen Fragen.

2.4 Eindeutigkeit des Informationsmaßes

Es läßt sich nun zeigen, daß eine Funktion H, die den folgenden Bedingungen genügt, bis auf eine multiplikative Konstante mit der oben definierten Informationsbedarfsfunktion

$$H(f) = -\sum_{i=1}^{n} p_f(a_i) \log p_f(a_i) = H(p_f(a_1), \ldots, p_f(a_n))$$

übereinstimmt:

1.
$$H(p_1, \ldots, p_n) \text{ ist stetig}$$

2.
$$H(p_1, \ldots, p_n) \leq H(\underbrace{\frac{1}{n},\ldots,\frac{1}{n}}_{n-mal})$$

3.
$$\begin{aligned} H(p_1, \ldots, p_n) &= H(p_1, \ldots, p_{i-1}, p_i + p_{i+1}, p_{i+2}, \ldots, p_n) \\ &+ (p_i + p_{i+1}) H\left(\frac{p_i}{p_i + p_{i+1}}, \frac{p_{i+1}}{p_i + p_{i+1}}\right) \end{aligned}$$

Prinzipiell würden diese drei Bedingungen schon hinreichen, um die Informationsfunktion 2.2 bis auf eine Konstante eindeutig zu definieren, der Beweis hierfür ist allerdings umfangreich ([Sh]), so daß ich noch eine weitere Bedingung mit aufnehme:

4.

$$H\Big(\underbrace{\frac{1}{nl},\frac{1}{nl},\ldots,\frac{1}{nl}}_{nl-mal}\Big) = H\Big(\underbrace{\frac{1}{n},\ldots,\frac{1}{n}}_{n-mal}\Big) + H\Big(\underbrace{\frac{1}{l},\ldots,\frac{1}{l}}_{l-mal}\Big)$$

Zum **Beweis** beschränken wir zunächst auf die Funktion

$$H^*(n) = H(\underbrace{\frac{1}{n},\ldots,\frac{1}{n}}_{n-mal})$$

Die Funktion $H^*(n)$ wächst monoton mit n, denn

$$\begin{aligned} H^*(n+1) &\geq H\Big(\frac{1}{2n},\frac{1}{2n},\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{(n-1)-mal}\Big) \\ &= H\Big(\underbrace{\frac{1}{n},\cdots,\frac{1}{n}}_{n-mal}\Big) + \frac{1}{n}H\left(\frac{1}{2},\frac{1}{2}\right) \\ &\geq H^*(n) \text{ weil } H\left(\frac{1}{2},\frac{1}{2}\right) \geq 0 \end{aligned}$$

Nun gilt weiter

$$H^*\left(n^{k+1}\right) = H^*\left(n^k \cdot n\right) = H^*\left(n^k\right) + H^*(n)$$

Daraus folgt durch mehrfaches Ausklammern

$$H^*\left(n^k\right) = k \cdot H^*(n)\ (\forall k \in \mathbb{N}).$$

Nun wählen wir ein $m \in \mathbb{N}$ derart, daß

$$2^m \leq n^k < 2^{m+1}.$$

also auch

$$m \leq \log_2(n^k) < m+1$$

oder

$$m \leq k \cdot \log_2 n < m + 1$$

Dann ist

$$H^*\left(2^m\right) \leq H^*\left(n^k\right) \leq H^*\left(2^{m+1}\right),$$

da H^* monoton wachsend ist. Mithin ist

$$m \cdot H^*(2) \leq k \cdot H^*(n) \leq (m+1) \cdot H^*(2)$$

und mit obiger Beziehung

$$\begin{aligned} (k \log_2 n - 1) \cdot H^*(2) \leq k \cdot H^*(n) &\leq (k \log_2 n + 1) \cdot H^*(2) \\ \left(\log_2 n - \frac{1}{k}\right) \cdot H^*(2) \leq H^*(n) &\leq (\log_2 n + \frac{1}{k}) \cdot H^*(2) \end{aligned}$$

Dies gilt für alle $k \in \mathbb{N}$, so daß im Grenzfalle $k \to \infty$ folgt:

$$\log_2 n \cdot H^*(2) \leq H^*(n) \leq \log_2 n \cdot H^*(2)$$

also ist für die spezielle Funktion $H^*(n)$ gezeigt, daß

$$H^*(n) = \log_2 n \cdot H^*(2) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \left(\frac{1}{n}\right) \cdot H^*(2)$$

Die multiplikative Konstante ergibt sich aus der Wahlmöglichkeit der Basis der Logarithmusfunktion (> 1).

Mit Hilfe der Eigenschaft (3) kann nun gezeigt werden, daß bei ungleichen rationalen Argumenten der Funktionsausdruck $H(p_1, p_2, \ldots, p_n)$ sich so umformen läßt, daß die Summanden nur Terme der Form $H^*(\cdots)$ sind.
Wir zeigen dafür zunächst die Bestimmbarkeit von $H\left(\frac{m}{n}, \frac{n-m}{n}\right)$ durch Induktion nach n und darauf aufbauend für jede beliebige Zahl von Argumenten.

Behauptung: Der Term $H\left(\frac{m}{n}, \frac{n-m}{n}\right)$ ist bestimmbar durch Terme der Form $H^*(\cdots)$, mit $n, m \in \mathbb{N}$, $\ltimes \geq \not\models$, $\not> < \ltimes$.

Induktionsanfang: $n = 2 \Rightarrow m = 1 \Rightarrow H\left(\frac{1}{2}, \frac{1}{2}\right) = H^*(2)$

Induktionsvoraussetzung: $H\left(\frac{m}{n}, \frac{n-m}{n}\right)$ sei bestimmbar für $n = 2, \ldots, k-1$;
$(m < n)$. Bleibt zu zeigen: $H\left(\frac{l}{k}, \frac{k-l}{k}\right)$ ist bestimmbar mit $l < k$.

$$\begin{aligned} H\left(\frac{l}{k}, \frac{k-l}{k}\right) &= H\left(\frac{1}{k}, \frac{l-1}{k}, \frac{k-l}{k}\right) - \frac{l}{k} \cdot H\left(\frac{1}{l}, \frac{l-1}{l}\right) \\ = \cdots &= H\Big(\underbrace{\frac{1}{k}, \ldots, \frac{1}{k}}_{l-mal}, \underbrace{\frac{1}{k}, \ldots, \frac{1}{k}}_{(k-l)-mal}\Big) - \sum_{j=2}^{l} \frac{j}{k} \cdot H\left(\frac{1}{j}, \frac{j-1}{j}\right) \\ & \quad - \sum_{j=2}^{k-l} \frac{j}{k} \cdot H\left(\frac{1}{j}, \frac{j-1}{j}\right) \end{aligned}$$

Die Terme der letzten Gleichung sind bestimmbar, da $H(\frac{1}{k}, \cdots, \frac{1}{k}) = H^*(k)$ ist, und in allen Fällen $j \leq l < k$ ist. Nach Induktionsvoraussetzung ist $H(\frac{1}{j}, \frac{j-1}{j})$ bestimmbar. Daher ist auch $H\left(\frac{l}{k}, \frac{k-l}{k}\right)$ bestimmbar.
Der vorliegende Induktionsbeweis läßt sich problemlos auf den Fall von mehr als zwei Argumenten übertragen. Für r Argumente $\frac{n_1}{n}, \cdots, \frac{n_r}{n}$ mit $\sum_{i=1}^{r} n_i = n$ ergibt sich im Induktionsschritt:

$$H\left(\frac{n_1}{n}, \cdots, \frac{n_r}{n}\right) = H\Big(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n_1-mal}, \cdots, \underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{(n_r)-mal}\Big) - \sum_{i=1}^{r} \sum_{j=2}^{n_i} \frac{j}{n} \cdot H\left(\frac{1}{j}, \frac{j-1}{j}\right)$$

Hieraus folgt die Bestimmbarkeit von H für beliebige rationale Argumente. Da H eine stetige Funktion sein soll und $\mathbb{Q}$ dicht in $\mathbb{R}$ liegt, ist die Bestimmbarkeit auch für beliebige reelle Argumente gezeigt.

■

2.5 Bedingte Fragen

Werden zwei unabhängige Fragen zu einer zusammengefaßt, so ist der Informationsbedarf der zusammengesetzten Frage gleich der Summe der Informationsnachfrage der einzelnen Fragen. Zwei Fragen f_1, f_2 sind unabhängig, wenn eine Antwort auf f_1 den Informationsbedarf von f_2 nicht beeinflußt und umgekehrt. Dies ist vergleichbar mit der Einführung unvereinbarer Ereignisse. Ist n_1 der Grad von f_1 und n_2 der von f_2, so ist $n_1 \cdot n_2$ der Grad der zusammengesetzten Frage $f_3 = f_1 \cdot f_2$. Eine Antwort auf f_3 setzt sich zusammen aus einer Antwort aus $\psi(f_1)$ sowie einer Antwort aus $\psi(f_2)$, die Antwortmenge $\psi(f_3)$ ist das Produkt dieser beiden Mengen: $\psi(f_3) = \psi(f_1) \times \psi(f_2) = \{a_i b_j | a_i \in \psi(f_1), b_j \in$

$\psi(f_2)\}$. Eine Antwort[8] auf f_3 kann nicht mehr Information enthalten, als die Summe der Informationen der Teilantworten.

$$\begin{aligned} H(f_3) &= -\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} p(a_ib_j)\log p(a_ib_j) \\ &= -\sum_{i=1}^{n_1} p(a_i)\sum_{j=1}^{n_2} p(b_j)(\log p(a_i)+\log p(b_j)) \\ &= -\sum_{i=1}^{n_1} p(a_i)\log p(a_i)\underbrace{\sum_{j=1}^{n_2} p(b_j)}_{=1} - \underbrace{\sum_{i=1}^{n_1} p(a_i)}_{=1}\sum_{j=1}^{n_2} p(b_j)\log p(b_j) \\ &= H(f_1)+H(f_2) \end{aligned} \tag{6}$$

Dieses Ergebnis stimmt mit der intuitiven Eigenschaft (3) in 2.2 überein, da die Fragen auch einzeln hätten gestellt werden können.

Ganz anders verhält es sich bei bedingten Fragen, etwa der Form „möchten Sie noch etwas trinken, wenn ja, was ?“ oder „welches ist das nächste gesendete Zeichen, wenn das aktuelle bekannt ist ?“
Zentraler Begriff ist hierbei der der bedingten Plausibilität, d.h. der Bereitschaft von S, auf das Eintreffen eines Ereignisses zu wetten, wenn es bestimmte Vorinformationen hat.

$$p(a_i|b_j) = \frac{p(a_ib_j)}{p(b_j)}$$

Das folgende Ergebnis ließe sich nun formal aus der Gleichung

$$H(f_1f_2) = -\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} p(a_ib_j)\log p(a_ib_j)$$

herleiten, indem für $p(a_ib_j)$ entsprechend $p(a_i|b_j)p(b_j)$ eingesetzt und geeignet umgeformt wird. Ich werde aber einen anderen Weg wählen, der die wichtigsten Etappen dieser Umformung als Definition einführt:

$$H(f_2|a_i) = -\sum_{j=1}^{n_2} p(b_j|a_i)\log p(b_j|a_i)$$

ist der Informationsbedarf der Frage f_2, wenn die Antwort a_i auf die Frage f_1 gegeben wurde.

[8] eine genauere Beschreibung des Antwortbegriffes erfolgt weiter unten.

$$H(f_2|f_1) = \sum_{i=1}^{n_1} p(a_i) H(f_2|a_i)$$

ist der erwartete Informationsbedarf von S, wenn die Antwort auf f_1 bekannt sein wird. Aus diesem Grund heißt dieser Ausdruck auch *Rückschlußinformation.*
Somit ergibt sich als Informationsbedarf der gemeinsamen Frage $f_1 f_2$:

$$H(f_1 f_2) = H(f_1) + H(f_2|f_1) \tag{7}$$

also gleich der Information, die S durch die Frage f_1 erlangen möchte plus der Information, die zusätzlich durch die Beantwortung der Frage f_2 erlangt wird, wenn die Antwort auf f_1 schon bekannt ist. Erfragen die Fragen f_1 und f_2 unabhängige Sachverhalte, so ist $H(f_2|f_1) = H(f_2)$ und die Gleichung 7 geht in die Gleichung 6 über.

2.6 Transinformation

Ein weiterer wichtiger Begriff zur Behandlung von Fragen ist der der Transinformation. S hat beim Stellen der Frage f_2 einen Informationsbedarf von $H(f_2)$ k-it, k hängt von der Basis des in H eingesetzten Logarithmus ab. Um an diese Information zu gelangen, kann S auch zuerst die Frage f_1 stellen, die über f_2 $H(f_2|f_1)$ k-it erfragt. Die Transinformation ist nun definiert als

$$I(f_1, f_2) = H(f_2) - H(f_2|f_1)$$

also den Informationsbedarf von f_2 minus dem Informationsbedarf über f_2, der übrigbleibt, wenn f_1 beantwortet ist. Anschaulich läßt sich dieser Wert interpretieren als die Menge an Information, die bei der Beantwortung von Frage f_1 über die Frage f_2 geliefert wird.
$I(f_1, f_2)$ hat nun einige interessante Eigenschaften, auf die im folgenden eingegangen wird:

1. $$I(f_1, f_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} p(a_i b_j) \log \frac{p(a_i b_j)}{p(a_i) p(b_j)}. \tag{8}$$

   Der Quotient $\frac{p(a_i b_j)}{p(a_i) p(b_j)}$ spiegelt dabei das angenommene Abhängigkeitsverhältnis zwischen a_i und b_j wider.

   Beweis:

$$
\begin{aligned}
I(f_1, f_2) &= H(f_1) - H(f_1|f_2) = H(f_1) - \sum_{j=1}^{n_1} p(b_j) H(f_1|b_j) \\
&= -\sum_{i=1}^{n_2} p(a_i) \log p(a_i) + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} p(b_j) p(a_i|b_j) \log p(a_i|b_j) \\
&= -\sum_{i=1}^{n_1} \left(\sum_{j=1}^{n_2} p(a_i b_j) \right) \log p(a_i) + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} p(a_i b_j) \log p(a_i|b_j) \\
&= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} p(a_i b_j) \left(\log p(a_i|b_j) - \log p(a_i) \right) \\
&= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} p(a_i b_j) \log \left(\frac{p(a_i b_j)}{p(b_j)} \frac{1}{p(a_i)} \right).
\end{aligned}
$$

2. $I(f_1, f_2) = I(f_2, f_1)$

   Diese folgt sofort aus Eigenschaft (1). Aus dieser Symmetrie resultiert auch der Name "Trans-Information", weil diese Information wechselseitig ist. Ist bekannt, daß Turm A um 10 Meter höher ist als Turm B, so liefert die Antwort auf die Frage f_1: „Wie hoch ist Turm A ?" auch die Antwort auf die Frage f_2 „Wie hoch ist Turm B ?", weil $H(f_2|f_1) = 0$ ist. Auf der anderen Seite liefert die Antwort auf f_2 auch die Antwort auf f_1, weil aus der Vorinformation bekannt ist, daß Turm B um 10 Meter kleiner ist als Turm A.

3. $H(f_1 f_2) = H(f_1) + H(f_2) - I(f_1, f_2)$

4. $0 \leq I(f_1, f_2) \leq H(f_2)$

Der Fall $I(f_1, f_2) = 0$ tritt ein, wenn die Fragen f_1 und f_2 voneinander unabängige Sachverhalte erfragen und somit die Antwort auf die eine nichts zur Beantwortung der anderen beisteuert. Und natürlich kann in f_2 nicht mehr Information über f_2 stecken als in f_2 selber. Da $H(f_2|f_2) = 0$ ist, ließe sich der Informationsbedarf von f_2 auch schreiben als $H(f_2) = I(f_2, f_2)$, also als die Information, die bei der Beantwortung von f_2 über f_2 erhalten wird.

Der Begriff der Transinformation wird noch mal im dritten Kapitel eine wichtige Rolle spielen, wenn es darum geht, Verhalten von Übertragungskanälen zu beschreiben.

2.7 Antworten

"Aufgabe von Wissenschaft ist es, Fragen in Fragen zu transformieren und zwar derart, daß die Fragen hinterher in irgendeiner Form "besser" sind als die Ausgangsfragen."
Diese zugegebenermaßen unexakte Beschreibung des wissenschaftlichen Handelns möchte ich dennoch meinem Konzept von Antwort zugrundelegen. Darüber hinaus ist es an der Zeit, die heuristische Beschreibung von Sachinformationsfragen etwas zu präzisieren: wichtig bei diesem Fragetypus ist die Menge der möglichen Antworten, die so gewählt werden muß, daß eine und nur eine Antwort zutrifft. Diese Menge sei im folgenden mit A bezeichnet. Für diese Menge ist es unerheblich, wie S eine konkrete Antwort einem Element a_i aus A zuordnet, ob über semantische Äquivalenz oder ähnliches. Eine Antwort auf die Frage „Gibt es im Kino heute eine Spätvorstellung ?" kann darin bestehen, daß der Fragesteller die Auskunft anruft, einen Bekannten fragt oder selber zum Kino geht um nachzuschauen. Wird im letzten Fall auch keine Sprache eingesetzt, so kann die Frage nach der Überprüfung vor Ort als beantwortet angesehen werden. Der Begriff "Antwort" bezeichnet hier also nicht die konkrete Sprechhandlung, sondern die Veränderung des Informationsstandes des fragestellenden Subjektes S. Von der Ursache dieser Veränderung wird in diesem Zusammenhang abgesehen. Eine Sachinformationsfrage f kann somit beschrieben werden als eine Menge $A = \{a_1, a_2, \ldots, a_n\}$ sowie einer Plausibilitätsverteilung ϕ auf A. $\Phi(n)$ bezeichne die Menge aller diskreten Plausibilitätsverteilungen auf einer n-elementigen Menge. Die Mächtigkeit von A, $|A| < \infty$, wird *grad* der Frage f genannt. Die Menge A heißt *Menge wesentlicher Antworten auf* f, ϕ heißt *Plausibilität der wesentlichen Antworten von* f.

Eine *Sachinformations-Antwort* α, im folgenden kurz *Antwort* genannt, ist eine Abbildung, $\alpha : \{A\} \times \Phi(n) \rightarrow \{A\} \times \Phi(n), \alpha(f) = \alpha((A, \phi)) = (A, \alpha^*(\phi))$, wobei $\alpha^* : \Phi(n) \rightarrow \Phi(n)$ die Plausibilitätsverteilung ϕ der Antworten auf f auf eine neue (nicht notwendig verschiedene) Plausibilitätsverteilung der wesentlichen Antworten auf f abbildet. Die Antwort α formt die Frage f in eine neue Frage $\alpha(f)$ um, die zwar die gleiche Menge wesentlicher Antworten -die Menge A- hat wie f, deren Plausibilitätsverteilung im allgemeinen aber eine andere sein wird. Im Idealfall ist dies eine Verteilung, die einer der möglichen Antwort den Wert 1, den übrigen die Plausibilität 0 zuordnet. Die Menge dieser ausgearteten Plausibilitätsverteilungen werde mit $\hat{\Phi}$ bezeichnet. Eine Frage $f = (A, \phi)$ heißt *durch die Antwort* α *vollständig beantwortet*, wenn $\alpha^*(\phi) \in \hat{\Phi}$ gilt. Eine Frage (A, ϕ), mit $\phi \in \hat{\Phi}$ wird *leere Frage* genannt, ihr Informationsbedarf ist nach Eigenschaft 2 der Informationsbedarfsfunktion identisch 0. Die Menge dieser Antworten, interpretiert als Abbildungen, sei mit $\mathcal{A}$ bezeichnet. Die Menge der wesentlichen Antworten sind solche Abbildungen aus $\mathcal{A}$, die die Frage f auf eine leere Frage abbilden. Insofern kann A als mit einer Teilmenge von $\mathcal{A}$ identifiziert werden.

Mit diesen Begriffen läßt sich nun auch der Wert angeben, den eine konkrete Antwort α für eine Frage f hat. Gemessen wird der Abstand des Informationsbedarfes der Frage f bezüglich der aus α resultierenden Folgefrage $\alpha(f)$.

$$I(\alpha, f) : \mathcal{A} \times \mathcal{F} \rightarrow \mathbb{R};\ \mathbb{I}(\alpha, \mathfrak{V}) = \mathbb{H}(\mathfrak{V}) - \mathbb{H}(\alpha(\mathfrak{V}))$$

Bezeichnet man die Folgefrage $\alpha(f)$ auch mit $f|\alpha$, d.h. die Frage, wie sie sich unter der Bedingung der Antwort a präsentiert, so lautet obige Formel

$$I(\alpha, f) = H(f) - H(f|\alpha)$$

Diese Formel weist optisch große Ähnlichkeiten zur Transinformation (2.6, S.29) auf, bezeichnet allerdings einen völlig unterschiedlichen Sachverhalt. Dort wurde der Zusammenhang zwischen zwei Fragen gemessen, hier der Zusammenhang zwischen einer Frage und einer Antwort. $I(\alpha, f)$ kann auch nicht symmetrisch sein, was allein schon an dem Definitionsbereich liegt.
Als grobe Abschätzung liegt der Wert einer Antwort zwischen völliger Desinformation und voller Beantwortung der Frage

$$-\log n \leq I(\alpha, f) \leq \log n$$

S kann zum Beispiel eine leere Frage stellen und durch eine Antwort vollständig verunsichert werden. Der Wert einer Antwort hängt ab von dem Informationsstand von S vor der Antwort. Dieser Informationsstand kann für eine genauere Abschätzung benutzt werden

$$-(\log n - H(f)) \leq I(\alpha, f) \leq H(f) \tag{9}$$

Eine Antwort kann also nicht mehr Information liefern, als gefragt worden ist und nicht mehr verwirren, als die Differenz zwischen dem Zustand der Desinformation zum Zeitpunkt des Fragens und der totalen Desinformation.

2.8 Fragestrategien

Selbst wenn eine Frage zur Zufriedenheit des Fragestellers beantwortet wird, können sich Folgefragen daran anschließen. Ist nun jeder Antwort einer Frage eine neue Frage zugeordnet, so kann bei mehrfacher Iteration ein Fragebaum entstehen, der in Abhängigkeit von den jeweiligen Antworten durchlaufen wird.

2.8.1 Exkurs: Bäume

Ein *Graph* ist ein Paar (E, K), wobei E die Menge der Ecken und $K \subseteq E \times E$ die Menge der je zwei Ecken verbindenden Kanten bezeichnet. In einem *gerichteten Graphen* haben die Kanten eine Orientierung, $k_1 = (e_1, e_2)$ ist eine

andere Kante als $k_2 = (e_2, e_1)$. Die Kante k_1 geht von der Ecke e_1 in Richtung der Ecke e_2, die Kante k_2 verläuft entgegengesetzt zu k_1. Eine Menge $\{(e_1, e_2), (e_2, e_3,), \ldots, (e_{n-1}, e_n)\}$ von gerichteten Kanten heißt *Pfad* von e_1 nach e_n. Bei einem Pfad ist die Orientierung aller Kanten die gleiche, andernfalls handelt es sich um einen *Weg*.
Ein Graph (E, K) heißt *Baum*, wenn zwei beliebige Kanten durch genau einen Weg verbunden sind. Ein Baum hat eine *Wurzel*, wenn alle Kanten von dieser weggerichtet sind, also von der Wurzel zu allen Ecken nur Pfade existieren. Die Knoten am Ende dieser Pfade heißen *Blätter*. Eine übliche Anschauung (und natürlich auch der historische Ursprung) eines Graphen ist in Form eines Diagramms gegeben, wo Ecken als Punkte und Kanten als Verbindungslinien zwischen diesen Punkten gezeichnet werden.

In unserem Fall ist die Wurzel die Ausgangsfrage f, die Kanten bezeichnen mögliche Antworten und die Ecken sind die Fragen, die sich aus der Beantwortung der ausgehenden Frage ergeben. Dies muß nicht zwangsläufig die oben definierte Folgefrage sein, d.h. in der Kante $\alpha_k = (f_i, f_j)$ bezeichnet f_j nicht notwendigerweise die Frage $\alpha_k(f_i)$. Die Frage f_j kann auch eine andere Frage bezeichnen, mit der S hofft, in seinem Sinne weiterzukommen. Ein Beispiel für einen Pfad in einem möglichen Fragebaum ist die Fragesequenz zwischen dem fragenden Subjekt S_1 und dem antwortenden Subjekt S_2:

S_1:„Was läuft heute abend in der Oper ?“ (A sei die Menge der Opern von Puccini, als Vorinformation von S_1, aus mangelndem Grund gleichgewichtet)
S_2:„Turandot oder Tosca.“(gleiche Gewichtung)
S_1:„Wann fängt es denn an ?“
S_2:„Sicherlich um 20:00 Uhr.“
S_1:„Wann fährt eine U-Bahn vom Hauptbahnhof bis zur Oper ?“
S_2:„Ich glaube, um 19:30 Uhr. Plus-Minus 5 Minuten.“
.....
Dieses Beispiel ist insofern auch instruktiv, als bei der letzten Frage auch Antworten wie “um 7:03“ möglich gewesen wäre, aus dem Kontext aber für S_2 ersichtlich war, daß S_1 pünktlich bei der Oper sein möchte. Als mögliche Antworten auf die Frage “Wann fährt eine U-Bahn vom Hauptbahnhof bis zur Oper?“ waren also nur Zeiten zwischen vielleicht 19:00 Uhr und 19:45 Uhr anvisiert. Jede andere Antwort hätte nicht nur das Kommunikationsspiel an dieser Stelle abgebrochen, sondern auch keine Information geliefert, eben weil für eine Antwort wie “um 7:03“ keine Frage gestellt worden war. Allerdings hatte die zweite Frage verschiedene Anfangszeiten in Erwägung gezogen, z.B. eine Matinee-Vorstellung um 11:00 Uhr am Sonntag oder eine Nachmittags-Vorstellung gegen 15:00 Uhr. Die Folgefrage klingt zwar in allen Fällen gleich, läßt dann aber je nach vorheriger Antwort nur Zeiten im Bereich von einer Stunde vor Vorstellungsbeginn als plausibel erscheinen. Insofern handelt es sich um einen im vornherein festlegbaren Fragebaum. Im Gegensatz dazu stehen Spiele, wie sie in der (zugegebener-

maßen vor langer Zeit abgesetzten) Sendung "Was bin ich ?" gespielt wurden: Ein Begriff, in diesem Falle ein Beruf, soll erraten werden, lediglich durch Stellen von binären Fragen, wobei negative Antworten sanktioniert werden. Dem Vorgehen der Fragesteller lag offensichtlich kein Fragebaum zu Grunde, sondern die Folgefragen wurden nach jeder Antwort in Abhängigkeit von dieser entwickelt. Ein Fragebaum wird im folgenden auch als *Fragestrategie* bezeichnet. Eine Fragestrategie $\mathcal{F}$ heißt *homogen*, wenn $f_1, f_2 \in \mathcal{F} \Rightarrow \psi(\{_\infty) = \psi(\{_\in)$, wenn also allen Fragen die gleiche Menge möglicher Antworten zugrundeliegt. Der Wert

$$\max_{f \in \mathcal{F}} grad(f)$$

wird *Grad der Fragestrategie* $\mathcal{F}$ genannt.

Der Begriff der Fragestrategie ist fundamental für den der Codierung im nächsten Kapitel.

3 Die Sätze von Shannon

In diesem Kapitel werden die Vorgänge der Nachrichtenübertragung, der Codierung und der Decodierung als Frage-Antwort-Vorgang interpretiert, womit sie sich zwanglos in die bisherige Begrifflichkeit einbetten lassen. Ziel ist die Herleitung der beiden Sätze von Shannon, die auch heute noch als Hauptsätze der Informationtheorie gelten.
Ausgehend vom Grundmodell der Kommunikation werde ich zuerst den Begriff der Quelle einführen, im Anschluß daran werden diese Quellen an einen Übertragungskanal angeschlossen. Falls der Kanal ungestört ist, erhält der Empfänger genau die Zeichen, die gesendet wurden und der Kanal kann somit vernachlässigt werden. Interessanter und realistischer ist der gestörte Kanal, der im zweiten Abschnitt behandelt wird.

3.1 Quellen

Eine Quelle wird interpretiert als stochastischer Prozess, d.h. als Folge von Zufallsvariablen. Je nachdem, ob die Folgeglieder unabhängig voneinander sind oder nicht, spricht man von gedächtnislosen Quellen oder von Quellen mit Gedächtnis, bei letzteren auch von Markov-Quellen. Beiden Quellen liegt ein endliches *Quellalphabet* zugrunde, also eine Menge $A = \{a_1, a_2, \ldots, a_n\}$, deren Elemente a_i *Zeichen* oder auch *Buchstaben* heißen. Das Subjekt S rechnet nun mehr oder minder stark mit dem Auftreten jedes Zeichens, d.h. für jeden Buchstaben a_i gibt es eine a-priori-Plausibilität p_i. Diese Plausibilität richtet sich z.B. nach der Struktur der bei der Übertragung benutzten Sprache und ist je nach Quelltyp von den vorangehenden Zeichen abhängig oder von ihnen unabhängig.

3.1.1 Gedächtnislose Quellen

Eine gedächtnislose *Quelle* $Q = (A, p)$ ist beschreibbar durch die Angabe eines Alphabetes $A = \{a_1, a_2, \ldots, a_n\}$ zuzüglich einer Plausibilitätsverteilung $p(A)$, die jedem Zeichen $a_i \in A$ seine Plausibilität p_i zuordnet. Hierbei bezeichnet n auch weiterhin die Mächtigkeit von A, also die Anzahl der Buchstaben des Quellalphabetes.
In den Begriffen des Kapitels 2 hieße dies, daß die Frage f:„Welches ist das nächste Zeichen, das von Q ausgegeben wird ?“, die Antwortenmenge
$A = \{$“Das Zeichen $'a_i'$“, $i = 1, \ldots n\}$ als wesentliche Antworten zuläßt, wobei der Antwort $'a_i'$ die Plausibilität p_i zugeschrieben wird. An dieser Stelle möchte ich auf eine semantische Feinheit hinweisen, die darin besteht, daß in der Antwort „Das Zeichen 'a_i'“ die Variable “a_i“ in Anführungszeichen steht. Natürlich lautet die Antwort, wenn a_i z.B. für den Buchstaben “B“ steht: „Das Zeichen “B““ und nicht: „Das Zeichen “a_i““. Auf der anderen Seite dürfen die Anführungszeichen auch nicht einfach weggelassen werden, weil die Aussage „Das Zeichen B“ eine Typverwirrung bedeuten und semantisch sehr unfein wäre. Deswegen benutze ich das Anführungszeichen '(dies hätte jetzt auch in Anführungszeichen stehen müssen, was der Lesbarkeit sicherlich nicht zuträglich gewesen wäre), um anzudeuten, daß die Zeichen zwischen diesen Anführungszeichen weiterhin als Variable zu deuten sind, die Anführungszeichen ““, wenn die so eingerahmten Zeichen wörtlich genannt sind. Ganze Sätze stehen, wie im Deutschen üblich, zwischen „“.
Der Informationsbedarf, der durch die Frage f zum Ausdruck kommt, ist

$$H_r(f) = H_r(Q) = H_r(p_1, \ldots, p_n),$$

wobei r die Basis der verwendeten Logarithmusfunktion angibt. Ist diese Basis nicht wichtig, so wird der Index weggelassen. S tut gut daran, die Plausibilitäten nach den tatsächlichen Verhältnissen der Quelle zu richten, um die Kohärenz in der Wahl der Plausibilitäten zu gewährleisten. Eine Verteilung, die den Auftrittsverhältnissen der Buchstaben der Quelle widerspricht, würde nämlich auf längere Sicht zu größeren Sanktionen (im Sinne des zweiten Kriteriums aus 1.3.1, S.7) führen, als notwendig wäre. Zur Vereinfachung wird der Informationsbedarf der Frage f in diesem Fall angesehen als eine Eigenschaft der Quelle[9] und kann so mit $H(Q)$ bezeichnet werden. In diesem Zusammenhang spreche ich dann von “Wahrscheinlichkeiten“ und “Plausibilitäten“, je nachdem, ob ich mehr die intersubjektive Eigenschaft der Quelle oder die Einschätzung des Subjektes betonen möchte.
$H(f) = H(Q)$ wird als *Zeicheninformation* von $Q = (A, p(A))$ bezeichnet. Im störungsfreien Fall, also dort, wo bei der Übertragung von Q zu S kein Fehler

[9] dies widerspricht durchaus nicht den methodischen Vorbemerkungen in Kapitel 2. Die statistische Eigenschaft der Quelle richtet sich natürlich auch hier nach der Erkenntnis des Subjektes, in diesem Fall allerdings unabhängig vom konkreten Subjekt, so daß vereinfacht von einer Eigenschaft der Quelle gesprochen wird.

auftritt, wird die Frage durch die Quelle vollständig beantwortet, indem ein Zeichen gesendet wird.
In vielen Fällen werden die von Q ausgehenden Zeichen zu Zeichenketten zusammengefaßt. Die *N-te Erweiterung der Quelle Q* faßt jeweils N Zeichen zu einer Kette zusammen. Mit $(A^N, p(A^N))$ sei die N-te Erweiterung der Quelle Q benannt. Werden die Zeichen unabhängig ausgesendet, was im Fall der gedächtnislosen Quelle immer der Fall ist, so ist die Plausibilität für das Auftreten der Kette

$$\zeta = q_{j_1} \cdots q_{j_N} \in Q^N : \; p(\zeta) = p_{j_1} \cdots p_{j_N}.$$

Hierbei gilt folgender

Satz 3.1 *Für die Zeicheninformation der N-ten Erweiterung Q^N einer Quelle Q gilt bei Unabhängigkeit der Quellzeichen in einer Kette:*

$$H(Q^N) = NH(Q)\square$$

Beweis: Sei $\zeta = q_{j_1} \cdots q_{j_N} \in Q^N$ eine beliebige Kette der Länge N, sozusagen ein Zeichen der N-ten Erweiterung. Dann gilt für die Zeicheninformation:

$$H(Q^N) = -\sum_{\zeta \in Q^N} p(\zeta) \log p(\zeta).$$

Aus der Unabhängigkeit der einzelnen Quellzeichen $q_{j_1} \cdots q_{j_N}$ von ζ folgt

$$\begin{aligned}
H(Q^N) &= -\sum_{i_1=1}^{n} \cdots \sum_{i_N=1}^{n} p(q_{i_1}) \cdot \ldots \cdot p(q_{i_N}) \log p(q_{i_1}) \cdot \ldots \cdot p(q_{i_N}) \\
&= -\sum_{i_1=1}^{n} \cdots \sum_{i_N=1}^{n} p(q_{i_1}) \cdot \ldots \cdot p(q_{i_N}) \left(\log p(q_{i_1}) \cdot \ldots \cdot p(q_{i_{N-1}}) + \log p(q_{i_N}) \right) \\
&= -\sum_{i_1=1}^{n} \cdots \sum_{i_{N-1}=1}^{n} p(q_{i_1}) \cdot \ldots \cdot p(q_{i_N}) \log p(q_{i_1}) \cdot \ldots \cdot p(q_{i_{N-1}}) \cdot \sum_{i_N=1}^{n} p(q_{i_N}) \\
&\quad -\sum_{i_1=1}^{n} \cdots \sum_{i_{N-1}=1}^{n} p(q_{i_1}) \cdot \ldots \cdot p(q_{i_{N-1}}) \sum_{i_N=1}^{n} p(q_{i_N}) \log p(q_{i_N}) \\
&= H(Q^{N-1}) \sum_{i_N=1}^{n} p(q_{i_N}) + \left(\sum_{i_1=1}^{n} p(q_{i_1}) \right) \cdot \ldots \cdot \left(\sum_{i_{N-1}=1}^{n} p(q_{i_{N-1}}) \right) H(Q) \\
&= H(Q^{N-1}) + H(Q).
\end{aligned}$$

Durch vollständige Induktion nach N wird der Satz bewiesen.

■

3.1.2 Markov-Quellen

Als nächstes betrachte ich Quellen, bei denen die Gedächtnislosigkeit und damit die Unabhängigkeit der einzelnen Zeichen nicht mehr gegeben ist. Eine solche Quelle wird *Markov-Quelle* genannt. Der Name rührt von folgender

Definition 3.1 *Eine Folge* $\{X_t, t \in \mathbb{N}\}$ *mit endlichem Wertebereich* $X_t \in \{a_1, a_2, \ldots, a_n\}$ *heißt* endliche Markov-Kette k-ter Ordnung*, wenn*

$$\begin{aligned} & \forall t > k \text{ und } \forall i, j_{k+1-t}, \ldots, j_k \in \{1, \ldots, n\} : \\ & p(X_t = a_i | X_0 = a_{j_{k+1-t}}, \cdots, X_{t-k} = a_{j_1}, \cdots, X_{t-1} = a_{j_k}) \\ = \quad & p(X_t = a_i | X_{t-k} = a_{j_1}, \cdots, X_{t-1} = a_{j_k}). \end{aligned}$$

Natürlich ist es auch interessant, die Zeicheninformation einer Markov-Quelle zu kennen. Ihre Bestimmung ist allerdings umständlicher und gelingt auch nur bei bestimmten, sogenannten ergodischen Quellen. Hierfür ist begriffliche Vorarbeit erforderlich:
Eine Markov-Kette k-ter Ordnung heißt *homogen*, wenn

$$\begin{aligned} \forall t > k : \qquad & p(X_t = a_i | X_{t-k} = a_{j_1}, \cdots, X_{t-1} = a_{j_k}) \\ = \quad & p(X_k = a_i | X_0 = a_{j_1}, \cdots, X_{k-1} = a_{j_k}), \end{aligned}$$

also wenn die beobachteten Abhängigkeiten unabhängig vom Beobachtungszeitpunkt sind.
Eine Markov-Kette heißt *stationär*, wenn

$$\forall i, j : \lim_{n \to \infty} p(X_{t+n} = a_i | X_t = a_j) = \lim_{n \to \infty} p(X_{t+n} = a_i) = p(a_i).$$

Dies besagt, daß die Plausibilität dafür, daß sich die Kette im Zustand a_i befindet, auf lange Sicht unabhängig ist von dem Ausgangszustand.
Eine Quelle mit dem Quellalphabet $A = \{a_1, a_2, \ldots, a_n\}$, die Zeichen mit der oben angeführten Eigenschaft einer Markov-Kette ausgibt, wird *homogene* bzw. *stationäre Markov-Quelle k-ter Ordnung* genannt.

Definition 3.2 *Jede Folge* $\zeta_j = a_{j_1} \cdots a_{j_k} \in A^k$ *von Buchstaben kann als* Zustand *einer homogenen Markov-Kette interpretiert werden, so daß der weitere Verlauf lediglich von dem aktuellen Zustand abhängt. Diese Abhängigkeit drückt sich in den* Übergangswahrscheinlichkeiten

$$p(a_i | \zeta_j) = p(X_t = a_i | X_{t-k} \cdots X_{t-1} = \zeta_j)$$

aus, die das nächste Ausgabezeichen a_i *und damit den Folgezustand* $\hat{\zeta}_j = a_{j_2} \cdots a_{j_k} a_i$ *in Abhängigkeit vom augenblicklichen Zustand bestimmen.*

Ein Zustand ζ_i heißt *rekurrent*, wenn er im Verlauf des Prozesses mit der Wahrscheinlichkeit 1 wieder auftritt. Geschieht dies im Mittel in endlichen vielen Schritten, so heißt ζ_i *positiv rekurrent*. Ist die Anzahl der Schritte zur Rückkehr nach ζ_i nur Vielfaches einer ganzen Zahl $m \geq 2$, dann heißt ζ_i *periodisch*, ansonsten *aperiodisch*. Kann jeder Zustand von jedem anderen aus erreicht werden, so heißt die Markov-Kette *irreduzibel*, ansonsten *reduzibel*. Eine irreduzibele Markov-Kette, mit positiv rekurrenten, aperiodischen Zuständen heißt *ergodisch*.
Die ergodischen Quellen sind besonders interessant, denn für sie gilt folgender, hier nicht bewiesener

Satz 3.2 (Ergodensatz) *Ist eine Quelle Q ergodisch, so existieren für alle Zustände stationäre Zustandswahrscheinlichkeiten*

$$p(\zeta_j) = p(a_{j_1} \cdots a_{j_k}) = \lim_{t \to \infty} p(X_{t-k} = a_{j_1}, \cdots, X_{t-1} = a_{j_k})$$

□

Die Frage f, die an die Quelle gestellt wird und von der Quelle beantwortet wird lautet nun: „Welches Zeichen wird als nächstes ausgegeben, wenn die letzten k Zeichen '$a_1 \ldots a_k$' sind (wenn die letzten k Zeichen bekannt sind) ?“
Betrachtet man alle Zeichenketten $\zeta_i = a_{i_1} \cdots a_{i_k} \in Q^k$ als mögliche Zustände der Quelle Q, so läßt sich f auch formulieren als „in welchen Zustand geht Q, wenn der augenblickliche Zustand 'ζ_i' ist ?“. Die Überzeugung von S, daß die Quelle in den Zustand ζ_j übergeht, wenn sie vorher im Zustand ζ_i war, wird beschrieben durch die bedingte Plausibilität $p(\zeta_j|\zeta_i)$. Da sich die beiden Zustände aber nur um das zuletzt gesendete Zeichen, sagen wir a_l unterscheiden, d.h. $\zeta_j = a_{j_2} \cdots a_{j_k} a_l$ mit $\zeta_i = a_{i_1} \cdots a_{i_k}$, lässt sich die Übergangsplausibilität auch als $p(a_l|\zeta_i)$ schreiben. Die Frage „welches Zeichen wird als nächstes ausgegeben, wenn der augenblickliche Zustand der Quelle 'ζ_i' ist ?“ erfragt also den gleichen Sachverhalt und hat somit die gleiche Darstellung wie die Frage „welches ist das nächste ausgegebene Zeichen, wenn die letzten k Zeichen '$a_1 \ldots a_k$' sind ?“. Diese Frage wird wegen ihrer Abhängigkeit von dem Zustand ζ_i auch als $f|\zeta_i$ bezeichnet, wobei f weiterhin für die Frage „welches ist das nächste ausgegebene Zeichen ?“ steht.
Der Informationsbedarf dieser Frage enthält natürlich bedingte Plausibilitäten:

$$H(f|\zeta_i) = H(Q|\zeta_i) = -\sum_{j=1}^{n} p(a_j|\zeta_i) \log p(a_j|\zeta_i)$$

Die Zeicheninformation einer ergodischen Markov-Quelle k-ter Ordnung ergibt sich nun aus

$$H(Q) = \sum_{\zeta_j \in Q^k} p(\zeta_j) H(Q|\zeta_j)$$

□

3.2 Kanäle

Die Tatsache, daß bei der Übertragung von Zeichen ein Medium, ein Kanal vorhanden sein muß, in dem die Übertragung erfolgt, bemerken Sender und Empfänger eigentlich erst, wenn auf Grund dieses Mediums Störungen bei der Übertragung vorkommen. Dies bedeutet, daß am Kanalausgang ein anderes Zeichen ausgegeben wird, als beim Senden vorgesehen war. Der Zusammenhang zwischen Eingabezeichen und Ausgabezeichen ist quantitativ erfassbar und es können Wahrscheinlichkeiten angegeben werden, nach denen die gesendeten Zeichen verfälscht werden. Ein Kanal läßt sich somit beschreiben als

- eine Menge von Eingabezeichen $A_K = \{a_1, a_2, \ldots, a_n\}$, auch *Eingabealphabet des Kanals* genannt.
- eine Menge von Ausgabezeichen $B_K = \{b_1, b_2, \ldots, b_m\}$, *Ausgabealphabet des Kanals* genannt.
- eine Familie von Wahrscheinlichkeiten, $P = \{p_{ij}, \sum_j p_{ij} = 1; i = 1, \ldots, n; j = 1, \ldots, m\}$, wobei p_{ij} die Wahrscheinlichkeit dafür angibt, daß bei gesendetem $a_i \in A_K$ das Zeichen $b_j \in B_K$ empfangen wird. Diese Familie kann man bei nicht zu umfangreichen Ein- und Ausgabealphabeten sinnvollerweise als Matrix darstellen, die sogenannte *Kanalmatrix*:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nm} \end{pmatrix}$$

Den so gegebenen Kanal bezeichne ich mit

$$[A_K, P, B_K]$$

Ein Kanal mit binärem Ein- und Ausgabealphabet $A_K = B_K = \{0, 1\}$ und Kanalmatrix

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

mit p als *Einzelfehlerwahrscheinlichkeit* heißt *symmetrischer Binärkanal*. Hierbei wird angenommen, daß die Fehler unabhängig voneinander und unabhängig von der eingegebenen Folge auftreten.
Hier muß unterschieden werden zwischen dem Quellalphabet der Quelle und dem Eingabealphabet des Kanals. Bei vielen Kanälen ist es nicht möglich, alle n verschiedenen Zeichen des Quellalphabetes zu übertragen, weil das Eingabealphabet des Kanals kleiner ist als das Quellalphabet. Überträgt der Kanal z.B. nur die Signale “Strom an“ und “Strom aus“, so stehen lediglich zwei Zeichen

zur Verfügung. Das Quellalphabet muß nun in die Zeichen übersetzt werden, die sinnvoll übertragen werden können, im Falle des Binärkanals in die Zeichen 0 und 1. Diese Übersetzung heißt "Codierung". Durch den Vorgang der Codierung wird die Quelle sozusagen an den Kanal angeschlossen. Der Code, der die Zeichen des Quellalphabets in die des Kanal-Eingabealphabetes übersetzt, muß im störungsfreien Kanal so gewählt werden, daß im Mittel möglichst wenige Zeichen zu übertragen sind. Eine untere Schranke der mittleren Wortlänge gibt der *1.Satz von Shannon* an. Im Falle des gestörten Kanals kann die Codierung dazu genutzt werden, auftretende Übertragungsfehler auf der Empfängerseite nicht nur zu erkennen, sondern in einem beschränkten Maße auch zu korrigieren.

3.3 Codes

Sei $A = \{a_1, \ldots, a_n\}$ das Alphabet einer Quelle. Eine Fragestrategie $\mathcal{C}$ vom Grad r, deren Ziel es ist, das gesendete Zeichen zu erfahren wird als *Code* bezeichnet. Wird jeder Antwort auf eine Frage $f \in \mathcal{C}$ eineindeutig ein Element der Menge $C = \{c_1, \ldots, c_r\}$ zugeordnet, dann heißt C *Codealphabet*. Das Codealphabet wird mit dem Kanal-Eingabealphabet gleichgesetzt, d.h. nur solche Codes sind sinnvoll, die auch tatsächlich übertragen weden können. Jedes Zeichen des Quellalphabetes wird durch eine (nicht für alle Zeichen notwendig gleiche) Anzahl Fragen identifiziert, ihm wird also ein *Codewort*, welches die Antworten in Kurzform notiert, zugeordnet. Das erste Zeichen des Codeworts gibt Aufschluß über die Antwort auf die erste Frage, das zweite über die auf die zweite Frage usw.
Ein Code läßt sich auch identifizieren mit der Menge seiner Codewörter, welche direkt aus den Blättern mit den dazugehörigen Zweigen des Fragebaums der Fragestrategie ablesbar sind. Insofern ist die Fragestrategie $\mathcal{C}$ identifizierbar mit dem Code und wird im folgenden auch so bezeichnet werden.
Bei der *Decodierung* wird von dem Codewort auf das ursprüngliche Zeichen des Quellalphabetes zurückgeschlossen. Ein Code heißt *nicht singulär*, wenn alle Codewörter verschieden sind, andernfalls heißt er *singulär*. Ein Code heißt *eindeutig decodierbar*, wenn jede Kette von Codebuchstaben durch höchstens eine Aneinanderreihung von Codewörtern erzeugt werden kann. Ein eindeutig decodierbarer Code heißt *sofort decodierbar*, wenn zu jedem Zeitpunkt die Decodierung ohne Kenntnis der Restfolge möglich ist. Ein Code heit *Blockcode*, wenn alle Codewörter die gleiche Länge haben.

3.3.1 Eindeutig decodierbare Codes

Im folgenden interessieren nur noch die eindeutig decodierbaren Codes, hier speziell die sofort decodierbaren.
Zunächst führe ich eine wichtige Ungleichung ein, die eine notwendige und hinreichende Bedingung angibt für die Existenz eines sofort decodierbaren Codes

(SD-Code).
Sei $W_l = c_{i_1} \dots c_{i_l}$ ein Codewort. Dann ist jedes Wort $W_j = c_{i_1} \dots c_{i_j}$ mit $j \leq l$ ein *Präfix* von W_l. Ein Code ist sofort decodierbar, wenn kein Codewort Präfix eines anderen Codewortes ist.
Es ist immer möglich, einen sofort-decodierbaren Code über einem Alphabet zu konstruieren, das mindestens zwei Buchstaben hat. Strebt man eine möglichst geringe Länge der Codewörter an, so gibt folgende Ungleichung Auskunft darüber, ob zu diesen Wortlängen ein SD-Code existieren kann:

Satz 3.3 (Satz von Kraft) *Eine notwendige und hinreichende Bedingung für die Existenz eines SD-Codes mit n Codewörtern der Länge $l_1, \dots, l_n$ ist die Gültigkeit der Ungleichung*

$$\sum_{i=1}^{n} r^{-l_i} \leq 1 \tag{10}$$

Hierbei gibt r die Mächtigkeit des Codealphabetes an.

□

Beweis: Aus einem Codealphabet mit r Zeichen lassen sich r^l verschiedene Wörter der Länge l konstruieren. Jedes Codewort der Länge $l_i \leq l$ ist Präfix von genau r^{l-l_i} Wörtern der Länge l, da es ebensoviele Möglichkeiten gibt, das Wort durch Anfügen von $l - l_i$ Zeichen auf die Länge l fortzusetzen. In einem präfixfreien Code scheiden für jedes Codewort der Länge $l_i \leq l$ seine r^{l-l_i} Fortsetzungen auf die Länge l aus, und es bleiben $r^l - \sum_{l_i \leq l} r^{l-l_i}$ mögliche Wörter der Länge l, die kein Codewort als Präfix haben. Daraus folgt nun direkt:

$$\forall l : r^l - \sum_{l_i \leq l} r^{l-l_i} \geq 0 \Leftrightarrow \forall l : \sum_{l_i \leq l} r^{-l_i} \leq 1 \Leftrightarrow \sum_{i=1}^{n} r^{-l_i} \leq 1.$$

■

Diese Bedingung ist ebenso notwendig (vgl. [Tz, S.19]) wie hinreichend für die Existenz eindeutig decodierbarer Codes, so daß unter dem Aspekt der möglichst kurzen Wortlänge die eindeutig decodierbaren Codes die gleichen Einschränkungen haben wie die SD-Codes, in der Folge also auch nur noch die letzteren betrachtet werden.

3.3.2 Der 1. Satz von Shannon

Unter Hinblick auf die Annahme, daß jede Frage mit einem Kostenaufwand verbunden ist, ist S natürlich daran interessiert, die Zahl der Fragen möglichst gering zu halten. Dies wird erreicht, indem die Fragen so geschickt gestellt werden, daß die Quellbuchstaben, die am plausibelsten sind, auch am schnellsten

erfragt werden können. Eine weitere Möglichkeit besteht darin, mehrere Quellzeichen zusammenzufassen und somit das Quellalphabet faktisch zu vergrössern, den Fragenaufwand pro Zeichen aber geringer zu halten. Das Zusammenfassen von N Quellzeichen wird beschrieben durch die N-te Erweiterung der Quelle Q.

Sei $Q = (A, p)$ eine gedächtnislose Quelle mit n Quellzeichen, $\mathcal{C}$ der zugehörige Code mit Codealphabet $C = \{c_1, c_2, \ldots, c_r\}$, der jedem Quellzeichen q_i ein Codewort der Länge l_i zuordnet, d.h. es müssen l_i Fragen gestellt (und beantwortet) werden, um das Zeichen q_i zu identifizieren. Dann bezeichnet $\overline{L} = \sum_{i=1}^{n} p_i l_i$ die *mittlere Codewortlänge*. Die mittlere Codewortlänge gibt an, wieviele Fragen im Mittel gestellt werden müssen, um ein gesendetes Zeichen zu erkennen.
Ein eindeutig decodierbarer Code heißt minimal, wenn die mittlere Codewortlänge $\overline{L}$ minimal ist. Hierbei gilt folgender

Satz 3.4 *Die Zeicheninformation $H_r(Q)$ ist eine untere Schranke für die mittlere Codewortlänge eines eindeutig decodierbaren Codes:*

$$H_r(Q) \leq \overline{L} = \sum_{i=1}^{n} p_i l_i$$

□

Beweis: Aus der Beziehung 3, S.21 der Informationsbedarfsfunktion folgt mit

$$q_i = \frac{r^{-l_i}}{\sum_{j=1}^{n} r^{-l_j}}$$

$$\begin{aligned} H(Q) &= -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i = -\sum_{i=1}^{n} p_i \log \frac{r^{-l_i}}{\sum_{j=1}^{n} r^{-l_j}} \\ H(Q) &\leq \sum_{i=1}^{n} p_i l_i \log r + \left(\sum_{i=1}^{n} p_i \right) \log \sum_{i=1}^{n} r^{-l_i} \\ &\leq \log r \sum_{i=1}^{n} p_i l_i. \end{aligned}$$

Wobei $\log \sum_{i=1}^{n} r^{-l_i} \leq 0$ direkt aus der Ungleichung 10, S.41 folgt.
Wegen

$$\log x = \frac{\log_r x}{\log r}$$

gilt:

$$H_r(Q) \leq \overline{L} = \sum_{i=1}^{n} p_i l_i$$

Bei einem Code mit n Codewörtern heißt dies, daß S davon ausgehen kann, daß mindestens $H_r(Q)$ Fragen gestellt werden müssen, um zu erfahren, welches Zeichen gesendet wurde.

■

Mit Hilfe dieses Satzes läßt sich nun relativ einfach der Fundamentalsatz der Codierung ohne Störung beweisen:

Satz 3.5 (1. Satz von Shannon im ungestörten Fall) *Beim Erfragen der Zeichen eines Klartextes auf der Grundlage eines n-buchstabigen Quell-Alphabetes, die zu Gruppen von N Buchstaben zusammengefaßt werden, kann man, wenn man N hinreichend groß wählt, immer eine Fragestrategie der Ordnung r finden, so daß der Mittelwert der Anzahl von Fragen, die zur Identifizierung eines Zeichens dienen, beliebig nahe bei $\frac{H(Q)}{\log r}$ liegt:*

$$\lim_{N\to\infty} \frac{\overline{L}_N}{N} = H_r(Q)$$

Hierbei bezeichnet $\overline{L}_N$ die mittlere Codewortlänge der N-ten Erweiterung der Quelle Q.

□

Vorläufig wird noch von einem Binär-Code ausgegangen, d.h. es dürfen lediglich Fragen gestellt werden, die zwei Antworten zulassen.
Zum Beweis wird die Fragestrategie von Shannon-Fano benutzt, die auf folgender Idee basiert:
Um mit einer Fragestrategie der Ordnung zwei das gesendete Zeichen zu identifizieren, müssen die möglichen Zeichen in zwei Gruppen aufgeteilt werden und die Zugehörigkeit zu einer der beiden Gruppen bestimmt werden. Diese Gruppe wird wieder unterteilt und so lange weitergefragt, bis die letzte Gruppe aus nur einem Zeichen besteht. Der Informationsbedarf, der in einer binären Frage enthalten ist, ist maximal $1bit$. Die Fragestrategie von Shannon-Fano beruht darauf, die Gruppen so einzuteilen, daß mit einer Frage auch möglichst viel Information erfragt wird.

Die Zerlegung ist derart zu wählen, daß die Plausibilität zur Zugehörigkeit zu einer der Gruppen nahezu gleich ist. Die Gruppen der ersten Zerlegung haben jeweils die Plausibilität nahe $\frac{1}{2}$, die der zweiten nahe $\frac{1}{4}$ usw. Die Quellzeichen werden eine l-stellige Codebezeichnung haben, die nach l Zerlegungen einer Gruppe angehören, die aus nur einem Buchstaben besteht. Die Plausibilität ihres Auftretens ist also (mit $r = 2$)

$$p_i = \frac{1}{2^{l_i}},\ l_i = \log_2 \frac{1}{p_i} = -\frac{\log p_i}{\log 2}$$

Dies wird nur in einigen Fällen der Fall sein. Im allgemeinen wird der Quotient $\frac{\log p_i}{\log 2}$ keine ganze Zahl sein. Wegen der Zerlegung in Gruppen möglichst gleich großer Plausibilität werden die Längen bei dieser Fragestrategie nahe bei diesem Quotienten liegen. Mit l_i wird also diejenige ganze Zahl bezeichnet, die nicht kleiner als $-\frac{\log p_i}{\log 2}$ ist, also

$$-\frac{\log p_i}{\log 2} \leq l_i < -\frac{\log p_i}{\log 2} + 1.$$

oder auch

$$l_i = \left\lceil -\frac{\log p_i}{\log 2} \right\rceil = \lceil -\log_2 p_i \rceil$$

Diese Aussagen lassen sich problemlos auf den Fall eines r- buchstabigen Codealphabetes übertragen, die Einteilung muß hier dann in r Gruppen möglichst gleicher Plausibilität erfolgen.
Ein SD-Code mit dieser Vorgabe ist möglich, weil die Ungleichung 10, S.41 erfüllt ist:

$$\sum_{i=1}^{n} r^{-l_i} = \sum_{i=1}^{n} r^{-\lceil -\log_r p_i \rceil} \leq \sum_{i=1}^{n} r^{\log_r p_i} = \sum_{i=1}^{n} p_i = 1$$

Am Ende des Beweises werde ich noch einen konkreten Code mit diesen Vorgaben konstruieren. Mit obiger Existenzaussage läßt sich der Beweis unter Ausnutzung bisheriger Ergebnisse relativ elegant zu Ende führen, da zusätzlich folgende Ungleichung gilt:

$$H_r(Q) \leq \overline{L} = \sum_{i=1}^{n} p_i \lceil -\log_r p_i \rceil \ < NH_r(Q) + 1 = \sum_{i=1}^{n} p_i(-\log_r p_i + 1).$$

Da diese Ungleichung für gedächtnislose Quellen gilt, stimmt sie insbesondere auch für ihre N-te Erweiterung Q^N:

$$H_r(Q^N) = NH_r(Q) \leq \overline{L}_N < NH_r(Q) + 1 = H_r(Q^N) + 1.$$

Nach Division durch N erhält man die Aussage des Satzes.

■

Abschliessend noch die versprochene Codierung, hier für den Fall $r = 2$, bei der die Länge eines Codewortes des Quellzeichens q_i genau die Länge $\lceil \log_2 p_i \rceil$ annimmt. Zunächst ordnen wir alle Quellzeichen nach abnehmender Plausibilität $p_1 \geq p_2 \geq \cdots \geq p_n$. Da einige dieser Werte gleich groß sein können, reicht die Plausibilität nicht, um einen Buchstaben zu charakterisieren. Wir bilden daher die Summen $P_1 = 0, P_2 = p_1, P_3 = p_1 + p_2, \ldots, P_n = p_1 + \cdots + p_{n-1}$. Diese Summen sind allesamt verschieden, und die Zahlen $P_1, \ldots, P_n$ können als spezielles

Alphabet aufgefaßt werden, das dem Quellalphabet eindeutig entspricht. Damit reicht es aus, dieses neue Alphabet zu codieren. Dafür stellen wir jede Zahl P_i dar als Binärbruch der Form

$$P_i = \frac{a_1}{2} + \frac{a_2}{2^2} + \cdots + \frac{a_j}{2^j} + \cdots,$$

wobei alle a_j entweder 0 oder 1 sind. Hiermit wird jedem P_i eine unendliche Folge $a_1 a_2 a_3 \ldots a_n \ldots$ der Ziffern 0 und 1 zugeordnet. Alle diese Folgen sind, wie die P_i, voneinander verschieden. Auch können diese Folgen sich nicht erst in weit vom Anfang entfernten Ziffern unterscheiden, denn alle $P_{i+1}, P_{i+2}, \ldots, P_n$ unterscheiden sich von P_i wenigstens um den Summanden p_i, d.h. um nicht weniger als $\frac{1}{2^{l_i}}$. Die Zerlegungen aller dieser Zahlen in Binärbrüche werden sich also von der Zerlegung der Zahl P_i nicht später unterscheiden als im Glied $\frac{a_{l_i}}{2^{l_i}}$. Daraus folgt, daß sich alle Binärbrüche von $P_{i+1}, P_{i+2}, \ldots, P_n$ bereits in den ersten l_i Ziffern vom Binärbruch P_i unterscheiden. Behalten wir vom Bruch P_i nur die ersten l_i Ziffern, so werden alle Zahlenfolgen $a_1 a_2 \ldots a_{l_i}$ paarweise verschieden sein und keine ist Anfang einer anderen. Offensichtlich ist es ein SD-Code, dessen Wortlängen gerade den Ansprüchen genügen. Die Rückübersetzung dieser Zahlenfolge in eine Fragestrategie, also die konkrete Formulierung der Fragen soll an dieser Stelle nicht vorgenommen werden, auch nicht die Verallgemeinerung zu r Codezeichen.

Wenn der Sender über den Kanal in einer Zeiteinheit L Fragen der Ordnung r beantworten kann, was L Codesignalen aus einem r elementigen Codealphabet pro Zeiteinheit entspricht, so sagt Satz 3.5, daß die Übertragungsgeschwindigkeit den Betrag

$$v = \frac{L \log r}{H} \text{ Zeichen pro Zeiteinheit}$$

nicht übersteigen kann, es ist jedoch eine Übertragung mit einer v beliebig nahekommenden Geschwindigkeit möglich. Die Größe

$$K = L \log r,$$

die im Zähler steht, ist unabhängig von der Nachricht und hängt nur vom verwendeten Kanal ab. Sie stellt die größte Menge an Informationseinheiten dar, die in einer Zeiteinheit übertragen werden können. K heißt daher *Durchlaßkapazität* des ungestörten Kanals.
Mit dem ersten Satz von Shannon ist eine unteren Schranke für die mittlere Codewortlänge bekannt. Die mittlere Codewortlänge wird im allgemeinen aber größer sein als dieser Wert, d.h. es werden mehr Zeichen übertragen als eigentlich notwendig sind. Bei r verschiedenen Codezeichen beschreibt der Quotient

$$E = \frac{H_r(Q)}{\overline{L}}$$

die *Effizienz* des Codes ,

$$1 - E = 1 - \frac{H_r(Q)}{\overline{L}}$$

heißt *Redundanz* oder *Weitschweifigkeit* des Codes. Die Redundanz ist ein Maß dafür, inwieweit die gesendete Menge an Zeichen die wirklich erforderliche Menge übersteigt.
Ungestörte Kanäle kommen in der Realität allerdings nicht vor, weil irgendeine Form von Störung die Nachrichtenübertragung immer behindern kann, so daß die bisherigen Überlegungen im folgenden Abschnitt entsprechend verallgemeinert werden.

3.4 Der gestörte Kanal

Bis zum gegenwärtigen Zeitpunkt mußten wir uns noch keine Gedanken um den Kanal machen, über den die Signale gesendet wurden, weil sie ohne Störung vom Sender zum Empfänger gelangten. Diese im allgemeinen unzulässige Annahme soll nun aufgegeben, das Modell entsprechend erweitert werden. Dennoch nehmen wir weiterhin an, daß dem Empfänger das vom Sender benutzte Quell-Alphabet bekannt ist und er sich bemüht, die gesendeten Zeichen aus den empfangenen zu rekonstruieren. Der Sender übersetzt die Quellzeichen in Ketten von Codezeichen und überträgt diese über den Kanal. Die Übertragung eines Codezeichens kann interpretiert werden als eine Antwort auf die Frage „welches ist das nächste gesendete Zeichen ?“. Der Empfänger versucht, aus den auf seiner Seite angekommenen Zeichen das ursprüngliche Quellzeichen zurückzugewinnen. Dies ist dann die Antwort auf die eigentlich interessante Frage „welches ist das ausgegebene Quellzeichen ?“. Bei der Übertragung können die bei der Übertragung benutzten Zeichen verfälscht werden. Ein solches empfangenes Zeichen kann als unklare Antwort auf die Frage „welches ist das nächste gesendete Zeichen ?“ interpretiert werden. Während im störungsfreien Fall jedes Zeichen als Antwort nur noch die leere Frage zurückließ, so kann im Fall des gestörten Kanals die Folgefrage immer noch einen echt positiven Informationsbedarf haben, der allerdings geringer ist, als vor dem Empfang des Zeichens. Wäre die Unsicherheit über das gesendete Zeichen nach dem Empfang sogar noch vergrößert, so wäre der Kanal praktisch unbrauchbar. Die Verteilung vor dem eigentlichen Senden heißt *a-priori* Verteilung. Sie spiegelt den Informationsbedarf wider, den S vor dem gesendeten Zeichen hat. Diese Verteilung wird durch das gesendete Zeichen modifiziert, im Ideal-, d.h. störungsfreien Fall vereinigt ein einziges Zeichen die Wahrscheinlichkeit 1 auf sich. Ein Kanal mit gleichem Ein- und Ausgabealphabet - und nur dieser wird im weiteren interessieren - $A_K = B_K = \{a_1, \ldots, a_n\}$ lässt sich als Spezialfall von Abschnitt 3.2 beschreiben durch eine Kanalmatrix $P = (p_{ij}) = (p(a_j|a_i)), i, j = 1, \ldots, n$, wobei die die i-te Zeile $p_{i_1} \cdots p_{i_n}$ die Plausibilitäten dafür angibt, daß die Buchstaben $a_1, \ldots, a_n$

gesendet wurden, wenn Buchstabe a_i empfangen wurde. Die i-te Zeile repräsentiert somit die Folgefrage, die bei der Antwort "Zeichen a_i wurde gesendet" entsteht. Im störungsfreien Fall ist als Spezialfall $P = E$ die Einheitsmatrix. Die Ausgabezeichen sind im Grunde genommen Plausibilitätsverteilungen auf den Eingabezeichen. Insofern läßt sich das Ausgabealphabet darstellen als eine Menge $V = \{v_1, v_2, \ldots, v_n\}$ von Verteilungen. Die Mächtigkeit von V ist so groß, wie die Mächtigkeit der Menge der Zeichen, die bei der Ausgabe möglich sind. In dem vorliegenden Fall ist die Menge der Ausgabezeichen identisch mit der Menge der Eingabezeichen. Diese Verteilungen lassen sich sofort aus der Kanalmatrix bestimmen und werden im weiteren auch durch diese dargestellt werden.
Da im Fall konkreter Realisierung eines Kanals die Übertragungsfehler und die daraus resultierende Kanalmatrix durch äußere Bedingungen gestaltet werden, läßt sich der Begriff der bedingten Plausibilitäten auch durch den der bedingten Wahrscheinlichkeiten ersetzen, da es sich hierbei um eine Eigenschaft des Kanals handelt, die unabhängig vom jeweiligen betrachtenden Subjekt S ist. S tut sogar gut daran, seine Meinung diesen Wahrscheinlichkeiten anzupassen, weil das Risiko einer Fehlentscheidung so minimiert wird. Im folgenden wird also der Begriff "Wahrscheinlichkeit" synonym zu dem der "Plausibilität" gebraucht (vgl. die Bemerkung zu Wahrscheinlichkeiten und Plausibilitäten bei Quellen, S.35).
Sind die Wahrscheinlichkeiten bekannt, mit denen die Zeichen $a_i \in A_K$ gesendet werden, so spiegelt die Frage, welches Zeichen wirklich gesendet wurde, einen Informationsbedarf von $H(A_K)$ wider. Sie wird hier mit f_1, bezeichnet. Die Frage, die klären soll, welches Zeichen beim Empfänger angekommen ist, f_2, hängt ab vom Ausgang der Situation, die f_1 erfragt. Beschrieben wird dies durch die bedingten Wahrscheinlichkeiten bzw. Plausibilitäten. Die Transinformation dieser beiden Fragen ist nach Formel 2.6, S.29

$$I(f_1, f_2) = H(f_1) - H(f_1|f_2).$$

Sie gibt an, inwieweit aus der Antwort auf die Frage f_2 auf die Antwort der Frage f_1 geschlossen werden kann. Diese Transinformation ist nur dann gleich dem Informationsbedarf von f_1, wenn aus dem empfangenen Zeichen eindeutig geschlossen werden kann auf das gesendete, d.h., die Antwort auf f_2 läßt keine Frage offen bzgl. der Antwort auf f_1. Dies ist genau der störungsfreie Fall. Die Transinformation ist gleich Null, wenn das gesendete vom empfangenen Zeichen unabhängig ist, und somit die Antwort auf f_2 keinerlei Hilfe zur Beantwortung von f_1 geben kann (und umgekehrt). In Verallgemeinerung der Durchlaßkapazität im störungsfreien Fall kann man im Falle eines Kanals mit Störung die Kanalkapazität festlegen als das erreichbare Maximum, mit der von der Antwort auf f_2 auf die von f_1 geschlossen werden kann:

$$k = \max_{p(a_i)} I(f_1, f_2)$$

Um die Kanalkapazität auszunutzen, müssen die Sendewahrscheinlichkeiten entsprechend gewählt werden. Im symmetrischen, binären Fall erreicht diese Kapazität ihr Maximum, wenn beide Zeichen mit gleicher Wahrscheinlichkeit gesendet werden (vgl. [Tz, S.39]).
Die Kapazität beträgt dann mit $H(p) = H(p, 1-p)$

$$I(f_1, f_2) = 1 - H(p),$$

wobei p die Fehlerwahrscheinlichkeit angibt.

3.4.1 Die Sätze von Shannon

In einem symmetrischen Kanal mit Störungen sind Fehler nur dann erkennbar, wenn nicht alle möglichen Eingabewörter zur Übertragung genutzt werden. Der Sender muß also einen größeren Code wählen, als eigentlich erforderlich ist. Z.B. könnte er jedes Zeichen zweimal senden um einen Fehler auf diese Weise erkennbar machen, wenn die beiden empfangenen Zeichen sich unterscheiden. Dies wäre dann ein Ein-Fehler-erkennender-Code, wenn auch ein sehr schlechter, weil sehr viele Fragen beantwortet werden müssen und keine Fehlerkorrektur möglich ist. Aufgabe der *Codierungstheorie* ist es, Codes zu entwickeln, die Fehler nicht nur erkennen, sondern auch korrigieren können. In allen Fällen wird dabei die Redundanz ausgenutzt, indem mehr Zeichen gesendet werden, als eigentlich erforderlich wäre. An die eigentlichen Codewörter werden noch zusätzliche Zeichen angehängt, die natürlich auch die mittlere Codewortlänge vergrößern. In diesem längeren Code sind allerdings nur einige der möglichen Zeichenketten Codewörter, auf die bei einer fehlerhaften Übertragung zurückgeschlossen werden kann.
Binäre Blockcodes mit Wörtern der Länge n dürfen also nicht alle 2^n Wörter benutzten. Die Sätze von Shannon für gestörte Kanäle, hier speziell für symmetrische Binärkanäle formuliert, zeigt den Zusammenhang zwischen der Kanalkapazität und der Anzahl nutzbarer Codewörter und der damit erreichbaren Fehlerrate.

Satz 3.6 (1. Satz von Shannon im gestörten Fall) *Es werden Zeichen über einen symmetrischen Binärkanal mit Einzelfehlerwahrscheinlichkeit p und Kapazität $k = 1 - H(p, 1-p) > 0$ gesendet. Von $2^n n$-stelligen Eingabewörtern werden $M(n) < 2^n$ als Codewörter benutzt. Dann gibt es für jedes $\epsilon \in \mathbb{R}^+$ Blockcodes der Länge n mit*

$$M(n) = \lfloor 2^{n(k-\epsilon)} \rfloor$$

Codewörtern, deren Fehlerwahrscheinlichkeit $p_E(n)$ bei der Decodierung mit wachsendem n beliebig klein werden

$$\lim_{n \to \infty} p_E(n) = 0$$

□

Daß diese fast sichere Übertragung nicht auf Kosten der Übertragungsgeschwindigkeit geht, besagt der

Satz 3.7 (2. Satz von Shannon im gestörten Fall) *Werden über einen symmetrischen Binärkanal mit Einzelfehlerwahrscheinlichkeit p und Kapazität $k = 1-H(p,1-p) > 0$ Zeichen aus einer Quelle Q mit Zeicheninformation $H(Q) < k$ gesendet, so läßt sich eine Codierung finden, so daß die Übertragungsgeschwindigkeit beliebig nahe an $H(Q)$ herankommt.*

□

Die Aussagen dieser Sätze sind von fundamentaler Bedeutung für die Informationstheorie, bzw die Theorie der Nachrichtenübertragung, sagen sie doch aus, daß es möglich ist, einen Code zu finden, mit dessen Hilfe der Empfänger mit beliebig nahe an 1 liegender Plausibilität auf die gesendete Nachricht schließen kann, wobei die Übertragungsgeschwindigkeit beliebig nahe an die Produktionsgeschwindigkeit der Quelle herankommt. Beides gilt allerdings nur unter der Bedingung, daß der Informationsbedarf, der durch die Anfrage entsteht, welches das nächste Zeichen ist, nicht größer ist, als die Durchlaßkapazität des Kanals. Außerdem kann ein langer Verzug in der Decodierung entstehen, so daß zwar die Übertragungsgeschwindigkeit fast gleich der Produktionsgeschwindigkeit ist, die Differenz zwischen Senden und Empfangen u.U. recht groß werden kann. Desweiteren sind die Beweise der Shannonschen Sätze im gestörten Fall nicht konstruktiv, d.h. sie geben keine Anleitung zum Bau dieses Codes an, so wie es im ungestörten Fall möglich ist. Ein guter Teil der Forschung im Bereich der Codierungstheorie liegt also im Auffinden eben jener günstigen Übertragungscodes. Deren Grundlage, also die Aussage, daß es überhaupt lohnenswert ist, solche Codes zu suchen, sind die beiden Shannonschen Sätze, weil sie aussagen, daß der Kanal beliebig gestört sein kann und trotzdem eine fehlerfreie Übertragung möglich ist, deren Geschwindigkeit beliebig nahe an der Produktionsgeschwindigkeit liegt, wenn $H(Q) < k$ ist.
Beide Sätze sollen hier nicht bewiesen werden.

3.5 Entscheidungsprobleme

Im Hinblick auf das nächste Kapitel wird nun beschrieben werden, wie sich der Decodiervorgang als Entscheidungsproblem beschreiben läßt. Hierbei kann die Voraussetzung des symmetrischen Binärkanals fallengelassen werden.
Wird die Zeichenfrage über einen gestörten Kanal gesendet, so ist die gesendete Antwort i.allg. eine nichtausgeartete Wahrscheinlichkeitsverteilung, die den Empfänger in Ungewißheit über das gesendete Zeichen läßt. Die fehlende Information („Welches Zeichen wurde denn tatsächlich gesendet ?“) läßt sich aus

Zeitgründen i.allg. nicht mehr erfragen, so daß der Empfänger sich mit Hilfe dieser Wahrscheinlichkeitsverteilung für ein Zeichen entscheiden muß. Dies ist der Fall einer Entscheidung bei Risiko, deren allgemeinere Charakteristik im nächsten Kapitel erörtert wird. Das empfangende Subjekt braucht eine Entscheidungsregel, die ihm angibt, welches Zeichen es aus den in Frage kommenden auszuwählen hat. Diese Entscheidung liegt noch vor dem Prozess der Decodierung, in den auch die Eigenschaften zur Fehlerkorrektur eingehen. Die Zeichen, auf die hier geschlossen werden soll, sind die kanalspezifischen. Natürlich können bei dieser Entscheidung auch Fehler auftreten. In diesem Fall muß dann die durch die Redundanz eingeführte Fehlerkorrekturmöglichkeit genutzt werden, um auf das tatsächliche Codewort zu schliessen. Genauer:

Definition 3.3 *Gegeben sei ein Kanal mit Eingabealphabet* $A_K = \{a_1, \ldots, a_n\}$ *und Menge der Ausgabeverteilungen* $V = \{v_1, v_2, \ldots, v_n\}$. *Eine* Entscheidungsregel *ordnet jeder Ausgabeverteilung* $v_i \in V$, *eindeutig ein Eingabezeichen* $\eta(a_j) \in A_K$ *zu.*

Definition 3.4 *Für eine Entscheidungsregel sei* $p(F|v_j)$ *die Wahrscheinlichkeit für eine falsche Zuordnung eines Eingabezeichens zu einer empfangenen Verteilung* v_j. *Hierbei bezeichnet* p_F *die Fehlerwahrscheinlichkeit einer Entscheidungsregel.*

Es gilt:

$$p(F|v_j) = 1 - p(\eta(v_j)|v_j) \text{ sowie } p_F = \sum_{j=1}^{n} p(v_j)p(F|v_j)$$

Die Fehlerwahrscheinlichkeit p_F bei gegebener Entscheidungsregel ist:

$$\begin{aligned} p_F &= \sum_{j=1}^{n} p(v_j)p(F|v_j) \\ &= \sum_{j=1}^{n} p(v_j)(1 - p(\eta(v_j)|v_j)) \\ &= 1 - \sum_{j=1}^{n} p(v_j)p(\eta(v_j)|v_j) \\ &= 1 - \sum_{j=1}^{n} p(\eta(v_j))p(v_j|\eta(v_j)) \end{aligned} \tag{11}$$

Sind die a-priori-Wahrscheinlichkeiten gleich, so gilt $p(\eta(v_j)) = \frac{1}{n}, j = 1, \ldots, n$ und somit

$$p_F = 1 - \frac{1}{n} \sum_{j=1}^{n} p(v_j|\eta(v_j)).$$

Ziel ist es, eine Entscheidungsregel zu finden, für die die Fehlerwahrscheinlichkeit p_F minimal ist.

Definition 3.5 *Eine* optimale Entscheidungsregel *ordnet jeder Ausgabeverteilung v_i das oder ein Eingabezeichen $\eta_{opt}(v_i)$ zu, für welches die bedingte Wahrscheinlichkeit $p(\eta_{opt}(v_j)|v_i)$ maximal ist, d.h.*

$$\forall v_i \in V : \; p(\eta_{opt}(v_i)|v_i) = \max_{a_j} p(a_j|v_j)$$

Diese nicht unbedingt eindeutige Entscheidungsregel veranlaßt S, sich für das Zeichen zu entscheiden, welches in der gegebenen Ausgabeverteilung mit der größten Wahrscheinlichkeit auftritt. Aus der Beziehung 11 folgt sofort:

Satz 3.8 *Für eine optimale Entscheidungsregel ist die Fehlerwahrscheinlichkeit p_F minimal.*

□

An dieser Stelle wird deutlich, wieso der bisherige Aufwand einer Begriffsumbildung getrieben wurde. Die Shannonschen Sätze lassen sich formulieren auf der Grundlage der Fragen „welches ist das nächste gesendete Zeichen?“ Aus den Aussagen dieses Kapitels ergibt sich dann, daß in der klassischen Shannonschen Informationstheorie der Begriff der Information nur noch auf den gesendeten Zeichen basiert und keinen Inhalt berücksichtigt, für den diese Zeichen Zeichen sind. Das Maß der Information ist in diesem Fall unabhängig von jeglicher Semantik und mißt lediglich, wie gut die Frage nach dem gesendeten Zeichen beantwortet wird. Dies ist eine durchaus berechtigte Frage im Bereich der Nachrichtentechnik, im normalen Lebensalltag ist man aber vielmehr an der Information interessiert, für die eben diese Zeichen Träger sind, in gesprochener Sprache auch unabhängig von der gewählten Formulierung. So erwartet ein Subjekt S auf die Frage „Ist dies der Bus nach Trinidad ?“ nicht eine bestimmte Antwort, sondern ist mit einem „Ja“ ebenso zufrieden zu stellen, wie mit einem „Yes, Sir“, einem „Mais oui, monsieur“ ebenso wie mit einem Kopfnicken. Natürlich wäre seine Frage auch im Negativ-Fall beantwortet, S wäre genauso informiert wie im positiven Falle, nur vielleicht nicht so zufrieden.
In der klassischen Theorie trägt eine Zeichenkette keine Information, wenn das empfangende Subjekt S vor dem Senden der Zeichen weiß, welche Zeichen gesendet werden. Dies stimmt mit dem Alltagsverständnis überein, nicht aber der Umkehrschluss, daß dann Information vermittelt wird, wenn die Zeichenkette nicht vorhersehbar ist. Auf die Frage „stelle ich ihnen gerade eine Frage ?“ ist die Antwort natürlich „Ja“, so daß S keine Information erwartet (dies gilt natürlich auch für Prüfungen, in denen der Prüfer die Fragen nicht zur eigenen Informationsbereicherung stellt). Allerdings ist die vom Antworter gewählte Zeichenkette nicht beliebig vorhersehbar, so eine Antwort kann von „Natürlich“, über dem erwarteten „Ja“ bis hin zu „was für eine Frage! Sicher.“ reichen, Fremdsprachen

noch nicht eingeschlossen. All diese Zeichenketten drücken den gleichen Sachverhalt aus, Information in diesem Sinne ist also unabhängig von der gewählten Formulierung.
Diese Unabhängigkeit wurde erreicht mit Hilfe der Äquivalenzrelation "ist semantisch gleich zu", die in Kapitel 2 eingeführt wurde. Im nun folgenden Kapitel führe ich die dort begonnen Überlegungen fort, wie Informationsgewinnung mit Hilfe von Frage- und Antwortspielen möglich ist. Die klassische Informationstheorie ordnet sich dann zwanglos ein, wenn die gewünschte Information darin besteht, zu erfahren, welches das gesendete Zeichen ist.

4 Information und Entscheidung

In diesem Kapitel geht es um das Problem, die Antwort auf eine bestimmmte Frage möglichst kostengünstig zu bekommen. Um dieses Problem mathematisch mit den Werkzeugen der vorghergehenden Kapitel beschreiben zu können, werde ich es erst umgangssprachlich vorformulieren, und diese sprachliche Beschreibung schrittweise exakter fassen.

4.1 Heuristik

Es wird von folgendem Szenario ausgegangen:
Das Subjekt S hat eine Sachinformationsfrage f, deren Beantwortung für S einen bestimmten Wert $W(f)$ hat. Die Einheit dieses Wertes bleibt momentan noch unbestimmt, d.h. es kann sich um rein subjektiven Nutzen, als auch um monetäre Werte handeln. S ist bereit, für die Antwort dieser Frage einen bestimmten Preis zu entrichten, der unterhalb des Wertes $W(f)$ liegt. Zur Beantwortung der Frage stehen eine Menge $\mathcal{T}$ von Informationsträgern zur Auswahl. Der Preis, den S beim Stellen der Frage an einen Informationsträger $T \in \mathcal{T}$ entrichtet, wird mit $K(f, T)$ bezeichnet. Dieser Preis ist unabhängig davon, ob die Frage beantwortet wird oder nicht. Der Preis sollte aber die selbe Einheit haben, wie der Wert, den die Frage für S hat. Für jeden Informationsträger $T \in \mathcal{T}$ rechnet S mit einer bestimmten Informationsmenge, die zwischen totaler Desinformation und dem Informationsbedarf der Frage f liegt. Das Ziel ist es nun, aus der Menge $\mathcal{T}$ denjenigen Informationsträger auszuwählen, der bei möglichst geringen Kosten möglichst viel Information liefert, wobei nur im Idealfall die Frage vollständig beantwortet ist. Unter Umständen kann sich S aber auch mit einer unvollständig beantworteten Frage zufrieden geben, wenn z.B. die Kosten für die volle Beantwortung zu hoch sind, oder die Frage von keinem $T \in T$ vollständig beantwortet werden kann. In diesem Fall muß S eine Entscheidung bei Unsicherheit treffen, wobei diese Entscheidung natürlich auch darin bestehen kann, zu warten, bis die Menge $\mathcal{T}$ sich geändert hat. Dies ist z.B. der Fall, wenn der Informationsträger eine Auskunftsstelle ist, die zum Fragezeitpunkt geschlossen ist, am folgenden Tag aber wieder zur Verfügung stehen

wird. S muß sich so entscheiden, daß aus der Entscheidung die für S geringsten (subjektive) Kosten entstehen.
Selbst wenn die Frage nicht voll beantwortet werden kann, ist es möglich, daß durch die Antwort das Risiko einer Fehlentscheidung für S ausreichend reduziert wird. In diesem Fall wird S zwar zuviel bezahlen, aber in dem Gefühl handeln, das Bestmögliche getan zu haben. Das Risiko der Entscheidung ist umgekehrt proportional zum Informationsstand von S. Das Risiko zu bestimmen, das eine bestimmte Handlung mit sich bringt, ist Aufgabe der Entscheidungstheorie.
Als zusätzliche Verfeinerung soll auch zugelassen werden, daß die Menge $\mathcal{T}$ noch nicht bekannt zu sein braucht, so daß diese Menge u.U. zunächst bestimmt werden muß. Die hierbei auftretende Frage f_2 wird *Frage zweiter Ordnung* genannt, weil es sich um eine Frage über die Frage f handelt. f_2 könnte z.B. lauten „Wo finde ich jemanden, der mir die Antwort auf die Frage 'f' geben kann ?“. Allgemeiner hat eine Frage die Ordnung n, wenn sie die Menge $\mathcal{T}$ einer Frage der Ordnung $n-1$ erfragt. Sachinformationsfragen, die keine Informationsträger erfragen, haben die Ordnung 1.
Fragen höherer Ordnung können auch als Sachinformationsfragen interpretiert werden, wenn aus einer Auswahl möglicher Informationsträger die tatsächlichen Informationsträger bestimmt werden sollen. Der Einfachheit halber werde ich Fragen höherer Ordnung als solche behandeln.
Fragen der Ordnung ≥ 2 sind typisch z.B. in Kaufhäusern, wo bei der Suche nach einem bestimmten Artikel zunächst bei der Information die Abteilung, innerhalb dieser Abteilung der entsprechende Fachverkäufer und über diesen der Standort des gewünschten Produktes erfragt werden muß. In einer fremden Stadt muß u.U. erst noch die Telefonnummer des Touristenbüros erfragt werden, um von dort die Adresse des entsprechenden Kaufhauses zu erfahren. Die Summe der hierbei entstehenden Kosten für S sollte natürlich auch unter dem Nutzen der Antwort, ausgedrückt durch den $W(f)$, liegen. Deswegen werden Informationsstellen immer sehr zentral eingerichtet, z.B. in der Nähe des Einganges oder direkt im Stadtzentrum, um den Aufwand für die Fragesteller möglichst gering zu halten.
Zusäzlich läßt sich eine komplexe Frage f unter Umständen geschickt in Teilfragen $f_1, \ldots f_n$ zerlegen, deren logische Verkettung dann die Antwort auf die ursprüngliche Frage f liefert. Dies ist vor allem dann sinnvoll, wenn die Frage f so speziell ist, daß kein Informationsträger gefunden werden kann, der eine zufriedenstellende Antwort liefert. Der hierbei auftretende Zerlegungsprozeß ist seinem Wesen nach ein kreativer und somit vermutlich nicht formalisierbar, so daß eine mathematische Beschreibung auszuschließen ist, zumindest hier nicht versucht wird. Zusätzlich muß auch die zugrunde liegende Logik angegeben werden, mit deren Hilfe aus den Antworten auf $f_1, \ldots, f_n$ auf die Antwort auf f geschlossen werden kann.

4.1.1 Exkurs: Logik

Unter “Logik“ wird häufig die aristotelische Prädikatenlogik, bzw. ihre Erweiterung verstanden, wie sie in der “Principia Mathematica“ von Whitehead und Russel verwendet wurde. Ich fasse hier den Begriff “Logik“ allgemeiner auf als die Lehre von den Begriffen, den Urteilen und den Schlüssen. Von der Logik hängt also ab, wie das entsprechende Subjekt zu seinen Begriffen kommt - über Definitionen, Anschauung, u.ä. - was es als wahr empf indet, und wie es aus bestehenden Wahrheiten zu neuen Wahrheiten kommen kann. In diesem Sinne kann es neben der deduktiven Logik, die vom Allgemeinen auf das Besondere schließt - der klassischen Logik - auch eine induktive Logik als Schlußweise vom Besonderen auf das Allgemeine geben. Andere Logikformen sind die Dialektische Logik im Sinne Hegels, oder die Pragmatische Logik im Sinne von Peirce, bei der Sätze als wahr angesehen werden, an denen kein Mitglied der idealen Kommunikationsgesellschaft mehr zweifelt, wobei systematischer Zweifel descartscher Tradition ausgeschlossen wird.
Als rational angesehen wird allerdings die deduktive Logik, die auch dem folgenden Beispiel zugrundeliegt, das die Zerlegung einer Frage in kleinere Teilfragen demonstriert, deren Antworten wiederum zur Antwort der Ausgangsfrage zusammengesetzt werden. Die deduktive, klassische Logik beruht auf drei Axiomen, hier entnommen aus [Fr]:

1. **A = A (Satz von der Identität)**
2. **A = ¬¬A (Satz vom Widerspruch)**
3. **Jedes x ist entweder A oder ¬A (Satz vom ausgeschlossenen Dritten)**

Im übrigen ist die Äquivalenzrelation “ = “ natürlich reflexiv, transitiv und symmetrisch.

Beispiel: Das folgende Beispiel ist entnommen aus [PM, S.3].
Drei Freundinnen bekommen jede Woche verschieden hohe Taschengeldbeträge. Folgende Informationen sind zu erlangen: Welches Mädchen (Vor- und Nachname) erhält wieviel, und wofür wird das Geld am liebsten ausgegeben ? Die Vornamen der Mädchen sind Marion, Susi und Uschi, die Nachnamen Bauer, Müller und Weber, die Taschengeldbeträge 6,-DM, 7,-DM und 8,-DM, die möglichen Vorlieben sind entweder Kleider, Süßigkeiten oder Schallplatten.
Vor jeder methodischen Überlegung ist also eine Tabelle mit $(3 \times 4) \times (3 \times 4)$ binären Einträgen auszufüllen, indem alle $3 \times 4 = 12$ Attribute miteinander in Beziehung gesetzt werden. Jeder Eintrag kann den Wert wahr=1 und falsch=0 annehmen. So bedeutet z.B. eine 1 im Feld Marion/6,-DM, daß Marion 6,-DM Taschengeld bekommt, eine 0 im Feld Kleider/Weber, daß das Mädchen mit Nachnamen Weber nicht am liebsten Kleider kauft usw. Mechanisch ließe sich

dies durch das Stellen von 144 Fragen vom Grad 2 erreichen. Dieses Vorgehen wäre allerdings äußerst unökonomisch, weil zum einen diese Tabelle noch verkleinert werden kann, wenn in Betracht gezogen wird, daß jedes Attribut mit sich selbst identisch, und die Beziehung zweier Attribute symmetrisch ist. Die entsprechend vereinfachte Tabelle benötigt dann nur noch 54 Einträge. Zum anderen kann durch logisches Kombinieren die Zahl der Fragen wesentlich reduziert werden. Minimal wird die Anzahl der Fragen, wenn die Antworten günstig, d.h. positiv ausfallen. Aus einer positiven Antwort lassen sich mehr Folgerungen ziehen als aus einer negativen. Die hier vorgestellten Antworten kommen mit lediglich zwei positiven Aussagen aus und beschreiben somit eher den Fall von ungünstig gewählten Fragen.
Folgende Hinweise sind in dem Rätsel noch gegeben:

1. Marion, die gerne Süßigkeiten kauft, bekommt weniger Taschengeld, als das Mädchen mit dem Nachnamen Weber.

2. Susi bekommt 7 Mark in der Woche

3. Uschi gibt ihr Geld nicht für Schallplatten aus. Weder sie, noch das Mädchen mit dem Nachnamen Bauer bekommt pro Woche 6 Mark Taschengeld.

Aus diesen Hinweisen lassen sich direkt neun Einträge ableiten, die auch als einzelne Fragen hätten gestellt werden können. Aus diesen neun Einträgen lassen sich dann durch logische Schlussfolgerungen alle übrigen Einträge ableiten (Übung). Wird davon ausgegangen, daß alle Fragen gleichviele Kosten verursachen, so wurden die Kosten für die vollständige Information auf ein sechzehntel des ursprünglichen Betrags reduziert.
Eine weitere Betrachtung, bzw. der Versuch einer Formalisierung soll hier nicht unternommen werden, ich wollte lediglich auf eine derartige Möglichkeit der Kostenminderung bei Fragen hinweisen, die im Normalfall auch praktiziert wird, wenn auch nicht immer mit dem Erfolg, wie im obigen Beispiel.

4.1.2 Algorithmus

Die heuristische Form des Algorithmus zur Informationsbeschaffung sieht bei diesem Stand so aus:

Dieser Algorithmus terminiert in endlicher Zeit, weil bei jedem Schritt entweder Nutzen aufgewendet oder das Vorgehen beendet wird. Ich gehe dabei davon aus, daß der Nutzen sich nicht asymptotisch der Null nähert, sondern immer größer oder gleich einem Wert $c > 0$ bleibt. Da der Nutzen, der aufgewendet werden kann, durch $W(f)$ beschränkt ist, wird S in endlicher Zeit eine Entscheidung fällen müssen, entweder bei Sicherheit oder bei Restrisiko. Die zu verfeinernden Begriffe bestimmen im wesentlichen den weiteren Aufbau dieses Kapitels, das sich wie folgt gliedert:
Zuerst gebe ich eine Einführung in die Entscheidungstheorie, wobei ich mich hierbei vor allem an dem Buch von Schneeweiß, [Sc], orientiere. Mit Hilfe der

in dieser Theorie eingeführten Begriffe wird es möglich sein, den Wert einer Frage näher zu bestimmen und die Abschätzung vorzunehmen, ob eine Antwort kostengünstig ist.

4.2 Entscheidungstheorie

Die klassische Entscheidungstheorie beruht im wesentlichen auf der Idee, daß ein entscheidendes Subjekt S aus mehreren Handlungsmöglichkeiten eine und nur eine Handlung auswählen muß. Die Umwelt befindet sich in einem von mehreren Zuständen, S weiß zum Zeitpunkt der Handlung allerdings nicht, welcher Zustand vorliegt. Die Menge dieser Welt- oder Umweltzustände heißt *Zustandsraum*. In der Begrifflichkeit des ersten Kapitels sind verschiedene Zustände unvereinbar, der Zustandsraum ist eine Einteilung. Jede Handlung führt in Kombination mit einem Umweltzustand zu bestimmten Konsequenzen, die in Form einer Gewinn/Verlust-Funktion ausgedrückt werden können. Ziel der klassischen Entscheidungstheorie ist es, eine Handlung so auszuwählen, daß diese Funktion maximiert wird. Sind S für die verschiedenen Umweltzustände Plausibilitäten ihres Eintretens bekannt, so ist eine strukturelle Ähnlichkeit mit dem oben eingeführten Konzept der Sachinformationsfrage offensichtlich. Hier entscheidet sich S für eine Handlung, dort für eine Antwort. Diese Antwort beschreibt einen Zustand der Welt. Wird davon ausgegangen, daß das Fürwahrhalten einer konkreten Antwort a_i, also die Annahme, der Zustand a_i treffe zu, als Handlung aufgefaßt werden kann, so liegt das klassische Entscheidungsproblem vor. Diese Ähnlichkeit führt dazu, daß die Begriffe der Informationstheorie, wie sie hier vorgestellt worden sind, zur Erweiterung des entscheidungstheoretischen Ansatzes benutzt werden können: Vor der eigentlichen Entscheidung für einen Weltzustand hat S noch die Möglichkeit, Informationen darüber einzuholen, welcher Zustand konkret vorliegt, mit dem Ziel, aus der Entscheidungssituation bei Unsicherheit eine Entscheidung bei Sicherheit herbeizuführen. Dies ist dann der Fall, wenn die Frage „welcher der folgenden Umweltzustände liegt vor: ... ?“ vollständig beantwortet wird. Aber auch bei unvollständiger Beantwortung läßt sich das Risiko einer Fehlentscheidung u.U. stark reduzieren. Diese Beziehung zwischen Informations- und Entscheidungstheorie soll im folgenden konkretisiert werden, zuvor noch eine kurze Abgrenzung von anderen Entscheidungstheorien: Ein Entscheidungsmodell, bei dem ebenfalls Informationen vor der eigentlichen Entscheidung eingeholt werden können, ist Gegenstand der statistischen Informationstheorie, vgl.[Ba][Di]. Ausgangslage dieser Theorie ist allerdings eine Ungewißheitssituation, d.h. eine Situation, in der keine Plausibilitäten für die möglichen Umweltzustände bekannt sind. Wesentlicher Inhalt der statistischen Entscheidungstheorie besteht in dem Auffinden von Plausibilitäten für die Umweltzustände mit Hilfe statistischer Verfahren. Das hierfür optimale Verfahren muß dazu ebenfalls in Abhängigkeit von dem jeweils vorliegenden Problem bestimmt werden. Sind alle Plausibilitäten bekannt, so reduziert sich das statistische Entscheidungsproblem auf das klassische. Im Gegensatz dazu kann S

bei dem obigen Algorithmus die bestehenden subjektiven Wahrscheinlichkeiten noch verändern und somit das Risiko einer Fehlentscheidung weiter minimieren.

4.2.1 Das Entscheidungsproblem

Es wird im folgenden davon ausgegangen, daß das Subjekt S mit der gestellten Informationsfrage ein bestimmtes Ziel verfolgt, daß z.B. aus der gewonnenen Antwort eine Handlung resultieren soll. S muß sich also für eine und nur eine Antwort entscheiden und wird somit zum *entscheidenden Subjekt*. Die Entscheidung wird getroffen zwischen den möglichen Antworten $a_1, \ldots, a_n$, die für S auf die Frage f plausibel erscheinen, den wesentlichen Antworten. Die Menge $\{a_1, \ldots, a_n\}$ wird *Entscheidungsraum* oder auch *Antwortenmenge* genannt und im weiteren mit $\mathcal{A}$ bezeichnet. Die Menge $\mathcal{A}$ wird über die Funktion $\psi(f)$, eingeführt in 2.1,auf Seite 19, bestimmt. Die Mächtigkeit dieser Menge sei n.
Die Suche nach der zutreffenden Antwort ist i.allg. kein Selbstzweck, sondern Grundlage für weitere geplante Handlungen. Die Menge der möglichen Handlungen wird mit $\mathcal{H}$ bezeichnet und es wird davon ausgegangen, daß jede Antwort genau eine Handlung zur Folge hat. Es kann weiter davon ausgegangen werden, daß S bei Kenntnis des vorliegenden Sachverhaltes die optimale Handlung ausführen wird, wobei der Begriff "optimal" noch zu präzisieren ist. Es existiert also eine Abbildung h aus der Menge der zur Auswahl stehenden Antworten $\mathcal{A}$ in die Menge der möglichen Handlungen H. Der Handlungsraum wird so eingeschränkt, daß h surjektiv ist. Da verschiedene Antworten die gleiche Handlung nach sich ziehen können, ist h i.allg. nicht injektiv, somit ist die Mächtigkeit von H kleiner oder gleich der Mächtigkeit von $\mathcal{A}$. Ergeben sich aus je zwei verschiedenen Antworten zwei unterschiedliche Handlungen, so sind die Mengen gleichmächtig.
Die Entscheidung für eine Antwort als Handlungsgrundlage richtet sich nach dem erwarteten Ergebnis der aus dieser Antwort resultierenden Handlung und dem damit verbundenen Gewinn. Die Entscheidungsituation besteht nun darin, daß S eine Antwort a^* aus $\mathcal{A}$ auswählen muß, ohne zu wissen, ob der mit a^* verbundene Sachverhalt tatsächlich besteht. S findet sich also mit einer von ihm unabhängigen Welt konfrontiert, deren konkreter Zustand S unbekannt ist. Diese Zustände sind gerade die Antworten, die S auf die Frage f als plausibel annimmt. Es wird angenommen, daß S alle für sein Problem relevanten Zustände der Welt kennt, allerdings nicht weiß, welcher gerade vorliegt. Die Menge der möglichen Weltzustände, der Zustandsraum, ist also identisch mit der Menge $\mathcal{A}$ der möglichen Antworten und wird im folgenden nicht mehr von dieser unterschieden, im Sinne der Vorbemerkungen in Kapitel 1. Entscheidet sich S für die Annahme, daß die Welt sich in dem Zustand befindet, der durch die Antwort a_i beschrieben wird, im folgenden nur noch der *Zustand* a_i genannt, so handelt es entsprechend $h(a_i)$. Befindet sich die Welt zum Zeitpunkt der Handlung im Zustand a_j, so entsteht für S aus der Handlung $h(a_i)$ im Zustand a_j ein Er-

gebnis aus einer noch nicht näher bestimmten Ergebnismenge $\mathcal{E}$. Es läßt sich also eine Funktion $e : \mathcal{A} \times \mathcal{A} \to \mathcal{E}$ angeben, die der Auswahl eines Umweltzustandes $a_i \in A$ zusammen mit dem tatsächlichen Zustand $a_j \in A$ ein Ergebnis $e(a_i, a_j) \in \mathcal{E}$ zuordnet.
Damit überhaupt eine Entscheidung gefällt werden kann, muß S eine Präferenzordnung $\preceq$ auf der Gesamtheit aller Ergebnisse $\mathcal{E}$ haben. Das heißt:
von je zwei Ergebnissen weiß S, welches es dem anderen vorzieht, oder ob sie ihm indifferent sind:

- Für je zwei Ergebnisse $e_1, e_2 \in \mathcal{E}$ gilt entweder $e_1 \preceq e_2$ oder $e_2 \preceq e_1$. Gilt $e_1 \preceq e_2$ und $e_2 \preceq e_1$, so ist $e_1 \sim e_2$, d.h. S ist indifferent gegenüber den Ergebnissen e_1 und e_2.
- wird das Ergebnis e_1 dem Ergebnis e_2 nicht vorgezogen und dieses dem Ergebnis e_3 nicht vorgezogen, dann wird auch e_1 dem Ergebnis e_3 nicht vorgezogen.

Das Entscheidungsproblem ist vollständig beschrieben, wenn noch die Art der Unsicherheit bekannt ist. Hierbei gibt es:

- Die *Spielsituation*: Die Zustände der Welt sind Handlungsmöglichkeiten rationaler Gegenspieler.
- Die *Risikosituation*: Das Eintreffen der Zustände wird von S mit bestimmten Plausibilitäten geschätzt.
- Die *Ungewißheitssituation*: Für die Zustände der Welt sind die Wahrscheinlichkeiten unbekannt und der Aufwand, Plausibilitäten zu bestimmen ist zu groß.

  Der Begriff "Ungewißheitssituation" ist wohl zu unterscheiden von dem Begriff "Unsicherheitssituation". Dieser bezeichnet ganz allgemein eine Entscheidungssituation, in der S nicht vollständig über den Weltzustand informiert ist, jener eine spezielle Unsicherheitssituation. Alle hier vorgestellten Typen bezeichnen Unsicherheitssituationen

Eine Darstellung des bisher Erörterten findet sich in der *Ergebnismatrix*, in der die Entscheidungen gegen die Umweltsituationen eingetragen werden, hier für den Fall der Sachinformationsfrage aufgestellt, wo der Entscheidungsraum mit dem Zustandsraum übereinstimmt:

	p_1	p_2	$\cdots$	p_n
	a_1	a_2	$\cdots$	a_n
a_1	e_{11}	e_{12}	$\cdots$	e_{1n}
a_2	e_{21}	e_{22}	$\cdots$	e_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_n	e_{n1}	e_{n2}	$\cdots$	e_{nn}

Die Wahrscheinlichkeiten $p_1, \ldots, p_n$ gelten nur für den Risikofall. Der Eintrag e_{ij} beschreibt das Ergebnis, das entsteht, wenn S sich für die Antwort a_i entscheidet, der tatsächliche Zustand aber a_j war. Dieser Zustand tritt mit der Plausibilität p_j ein. Die Präferenzrelation auf $\mathcal{E}$ läßt sich mittels einer Nutzenfunktion u in der Matrix darstellen, indem die Ergebnisse e_{ij} durch ihren Nutzen $u(e_{ij}) =: u_{ij}$ ersetzt werden. So ensteht die *Entscheidungsmatrix* oder *Nutzenmatrix* $\mathcal{U}$, in der Spieltheorie auch Auszahlungsmatrix, in der statistischen Entscheidungstheorie auch Verlust- oder Gewinnmatrix genannt:

	p_1	p_2	$\cdots$	p_n
	a_1	a_2	$\cdots$	a_n
a_1	u_{11}	u_{12}	$\cdots$	u_{1n}
a_2	u_{21}	u_{22}	$\cdots$	u_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_n	u_{n1}	u_{n2}	$\cdots$	u_{nn}

In der Entscheidungsmatrix stehen also Zahlen, im Gegensatz zur Ergebnismatrix, deren Elemente von der jeweiligen Entscheidungssituation abhängen.

4.2.2 Entscheidungskriterien

Bei der Entscheidung für einen Zustand (und der damit verbundenen Handlung) kann sich S von Entscheidungskriterien leiten lassen. Ein Entscheidungskriterium, das zu jedem Entscheidungsproblem die optimalen Antworten auswählt (die dann untereinander indifferent sind), heißt *Entscheidungsregel*. Wird die Auswahl lediglich eingeschränkt, so spricht man von einem *Entscheidungsprinzip*. Die hier vorgestellten Entscheidungskriterien stellen eine Präferenzrelation im Bereich der Antworten auf. Diese Relation legt fest, ob von zwei Antworten eine als eher zutreffend angesehen wird oder ob beide als indifferent gesehen werden. Dies ist nicht identisch mit den Plausibilitäten für die Antworten, weil S nicht unbedingt die plausibelste Antwort als die zutreffende ansieht, z.B. weil mit einer unplausibelen Antwort ein großer Verlust verbunden ist, wenn sie zutrifft, S sich aber für eine andere entschieden hat. In diesem Fall kann S eine "worst case" Entscheidung treffen, um den möglichen Schaden gering zu halten. Empfindet S die Antwort a_1 als eher zutreffend als die Antwort a_2, so sagt man, S zieht die Antwort a_1 der Antwort a_2 vor und schreibt

$$a_1 \succ a_2 \text{ oder auch } a_2 \prec a_1$$

Die Indifferenz zwischen den Antworten a_1 und a_2 wird als

$$a_1 \sim a_2$$

geschrieben. Wird a_1 der Antwort a_2 vorgezogen oder sind beide Antworten indifferent, so schreibt man:

$$a_1 \succeq a_2.$$

Die Indifferenz von a_1 und a_2 besagt somit das gleiche, wie

$$a_1 \succeq a_2 \text{ und } a_2 \succeq a_1.$$

Eine Präferenzrelation auf der Menge $\mathcal{A}$ der Antworten ist nicht zu verwechseln mit einer Präferenzrelation auf der Menge $\mathcal{E}$ der Ergebnisse. Die letztere ist Teil des Entscheidungsproblems, während die erstere noch ermittelt werden soll. Mit Hilfe der Entscheidungskriterien soll aus der Präferenzstruktur der Ergebnismatrix, ausgedrückt durch die Nutzenfunktion u, auf die Präferenzstruktur der in Frage kommenden Antworten geschlossen werden.
Im folgenden werden einige Beispiele für Entscheidungsregeln gegeben. Die Vorgehensweise dieser Regeln ist immer die gleiche: es wird jeder Antwort ein von der jeweiligen Regel abhängiger Wert zugeordnet. Von zwei Antworten wird die mit dem höheren Wert vorgezogen. Antworten mit gleichem Wert sind zueinander indifferent.

(1) *Wald-Regel* (auch Minimax-Regel):

$$a_i \succeq a_j, \text{ wenn } \min_k u_{ik} \geq \min_k u_{jk}.$$

Die Wald-Regel bewertet alle Antworten nach ihren schlechtestmöglichen Konsequenzen. Sie eignet sich vorwiegend für die Spielsituation, wenn die Umwelt ein rational handelnder Gegenspieler ist. Bei einer normalen, dem Subjekt indifferent gegenüberstehenden Umwelt, ist solch ein Pessimismus unbegründet. Eine direkte Erweiterung ist daher die

(2) *Hurwicz-Regel* mit dem Optimismusparameter $\lambda,\ 0 \leq \lambda \leq 1$:

$$a_i \succeq a_j, \text{ wenn } (1-\lambda)\min_k u_{ik} + \lambda \max_k u_{ik} \geq (1-\lambda)\min_k u_{jk} + \lambda \max_k u_{jk}.$$

Die Hurwicz-Regel zieht neben dem schlechtesten auch das beste Ergebnis in Betracht. Beide Werte werden zu einem gewichteten Mittel zusammengefaßt, wobei die Gewichtung das mehr oder minder pessimistische Verhalten widerspiegelt. Je kleiner λ ausfällt, desto größer ist der Pessimismus von S. Für $\lambda = 0$ ergibt sich die Wald-Regel, im Falle $\lambda = 1$ ist die Erwartung absolut optimistisch auf das beste Ergebnis hin ausgerichtet.

(3) *Savage-Niehans-Regel*:

$$a_i \succeq a_j, \text{ wenn } \min_k (u_{ik} - \max_h u_{hk}) \geq \min_k (u_{jk} - \max_h u_{hk}).$$

Die Savage-Niehans-Regel ist auch unter dem Namen "minimal-regret" Regel bekannt. Der Ausdruck $u_{ik} - \max_h u_{hk}$ mißt den Unterschied zwischen dem Ergebnis bei Wahl von Antwort a_i und der bestmöglichen Entscheidung, wenn tatsächlich a_k der Fall ist. Dieser Wert kann auch als Bedauern (engl.: regret) aufgefaßt werden, von der Form: „ach, hätte ich mich doch nur für a_k entschieden". Dieses Bedauern versucht die Regel von vorneherein möglichst klein zu halten. Der größte Nachteil dieser Regel ist, daß sich die Präferenzstruktur auf dem Antwortenraum durch Hinzufügen oder Weglassen auch nicht-optimaler Antworten ändern kann.
Die ersten drei Regeln galten auch für den Ungewißheitsfall, weil sie ohne eine Plausibilitätssverteilung auf dem Antwortenraum auskommen, die nun folgende Bernoulli-Regel, sowie die Regel von Hodges und Lehmann gelten dahingegen nur für den Risikofall.

(4) *Bernoulli-Regel*:

$$a_i \succeq a_j, \text{wenn} \sum_k u_{ik} p_k \geq \sum_k u_{jk} p_k.$$

Diese Regel bewertet Antworten nach dem Erwartungswert ihres Nutzens. Die Antwort mit dem höheren Erwartungswert wird vorgezogen. Eine genauere Untersuchung dieser Regel erfolgt weiter unten. Die Bernoulli-Regel enthält als Spezialfall die

(5) *Laplace-Regel*:

$$a_i \succeq a_j, \text{wenn} \sum_k u_{ik} \geq \sum_k u_{jk}.$$

Die Laplace-Regel basiert auf dem Prinzip vom mangelnden Grunde. Hierbei werden alle Umweltzustände als gleichplausibel angesehen, weil es keinen Grund gibt, einen Zustand als plausibler anzunehmen als einen anderen.

(6) *Hodges-Lehmann-Regel* mit dem Vertrauensparameter $\lambda, 0 \leq \lambda \leq 1$:

$$a_i \succeq a_j, \text{ wenn } \lambda \sum_k u_{ik} p_k + (1-\lambda) \min_k u_{ik} \geq \lambda \sum_k u_{jk} p_k + (1-\lambda) \min_k u_{jk}.$$

Diese Regel ist wie die Regel von Hurwicz ein gewichtetes Mittel, hier aus der Bernoulli- und der Wald-Regel. Der Parameter λ gibt dabei das Vertrauen wieder, das S in seine subjektiven Wahrscheinlichkeiten setzt. Für $\lambda = 1$ ist das Vertrauen ungetrübt und die Bernoulli-Regel wird unverändert angewandt. Im

Falle $\lambda = 0$ werden die Plausibilitäten überhaupt nicht mehr berücksichtigt und die Entscheidung erfolgt nach der Minimax-Regel.
Selbstverständlich lassen sich noch andere, eventuell mehrfache Mischungen denken, wodurch die daraus resultierenden Regeln dem entsprechenden Problem und dem Charakter des Entscheidungsträgers angepaßt werden können.
Alle oben stehenden Kriterien legen für jedes Entscheidungsproblem eine eindeutige Reihenfolge der Antworten fest.
Unter diesen Regeln wird einzig und allein die Bernoulli-Regel als rationales Entscheidungskriterium gelten (s.u.). Die übrigen Regeln finden in der Praxis zwar Anwendung, müssen aber als irrational eingestuft werden, wie z.B. die Wald-Regel als maximal pessimistisch einzustufen ist, solange keine Spielsituation vorliegt. Dennoch dienen diese Regeln als Analogon zum Bestimmen des Wertes, den eine Frage für S hat.

4.2.3 Der Wert einer Frage

Im weiteren wird ausschließlich die Risikosituation behandelt, wegen ihrer formalen Ähnlichkeit zu Sachinformationsfragen. Die subjektiven Wahrscheinlichkeiten oder Plausibilitäten werden benötigt, um die Ergebnisse der Informationstheorie anwenden zu können. Sollte S keine Vorinformationen über die Umweltzustände haben, so gilt das Prinzip des mangelnden Grundes und die in Frage kommenden Antworten werden als gleichwahrscheinlich eingestuft. So kann gewährleistet werden, daß die sukzessive Verbesserung des Entscheidungsrisikos (s.u.) mit Hilfe des im Abschnitt 4.1.2 eingeführten Algorithmus möglich ist. Eine ausführliche Diskussion über subjektive Wahrscheinlichkeiten wurde im ersten Kapitel geführt.
An dieser Stelle muß näher präzisiert werden, wie die Einträge der Ergebnismatrix aussehen sollen und wie S für sich den Wert einer Frage festlegen kann.
Das Stellen einer Frage erfordert von S Zeit, Geld und psychologische Kosten, z.B. Aufwand, Überwinden von Hemmschwellen, usw. Je nach Informationsträger werden diese Kosten unterschiedlich hoch ausfallen. Bestimmt werden sie durch die drei Funktionen $Z, G, P : \mathcal{T} \to \mathbb{R}$. Handelt S in einer für S günstigen Sitation, so ist ein Gewinn von Geld und Zufriedenheit (psychologischer Nutzen) möglich. Eine Handlung kostet aber in jedem Fall Zeit. Die Einträge in der Ergebnismatrix sind somit Vektoren, in diesem Fall 3-dimensionale Vektoren mit den Dimensionen Zeit, Geld und Zufriedenheit. Vor allem bei letzterer ist sowohl die Angabe einer Einheit, als auch die Messung eher intuitiv als konkret meßbar einzustufen. Am ehesten ist eine Umrechnung in Geld vorstellbar, wenn der Zuwachs von Geld ebenfalls mit einem Zuwachs an Zufriedenheit verknüpft ist. Die drei Dimensionen werden je nach Situation von S unterschiedlich gewichtet. Als Nutzenfunktion u läßt sich nun die gewichtete Summe dieser Vektorkomponenten einführen, die den Ergebnisvektor in eine reelle Zahl umwandelt. Diese

Werte haben den Vorteil, direkt miteinander vergleichbar zu sein, sie sind kardinal. Kardinaler Nutzen ist bis auf die Nutzeneinheit und den Nutzennullpunkt eindeutig festgelegt. Die Einheit spielt keine wesentliche Rolle und wird im folgenden als NE (Nutzeneinheit) bezeichnet. Ein Nullpunkt kann dann festgelegt werden, wenn das Ergebnis als Veränderung der augenblicklichen Situation verstanden wird. Diese ist dann der Nullpunkt. Wie oben erwähnt, wird S bei Kenntnis der zutreffenden Antwort eine optimale Handlung durchführen, d.h. die, die den meisten Gewinn verspricht. Die Entscheidungsmatrix wird also so aussehen, daß

$$u_{ii} \geq u_{ji} \quad \forall i, j = 1, \ldots, n = |A|$$

gilt. Gleichheit kann auftreten, wenn unterschiedliche Antworten gleiche oder gleichwertige Handlungen zur Folge haben.
Damit ist es nun möglich, den Wert $W(f)$ einer Frage f anzugeben, wobei es hier, wie bei den oben angeführten Entscheidungsregeln, verschiedene Möglichkeiten gibt, diesen Wert zu bestimmen. Ihre jeweilige Anwendung hängt ab von der Frage f und dem handelnden und entscheidenden Subjekt S. Vorausgesetzt ist auch hier wieder die Liste der möglichen Antworten, sowie die vollständige Nutzenmatrix.

(1) *Wald-Wert* (auch Minimax-Wert):

$$W(f) = \min_j \max_i u_{ij} = \min_i u_{ii}$$

Hier ist S nicht bereit, mehr für eine Antwort von f auszugeben, als der kleinste sichere Gewinn. S geht davon aus, daß die Antwort richtig ist, die den kleinsten aller Gewinne realisieren läßt, selbst bei optimaler Entscheidung. Hierin drückt sich der gleiche Pessimismus aus, wie in der Minimax-Entscheidungsregel. Ebenso gibt es hier auch den

(2) *Hurwicz-Wert* mit dem Optimismusparameter $\lambda, 0 \leq \lambda \leq 1$:

$$W(f) = (1 - \lambda) \min_j \max_i u_{ij} + \lambda \max_i \max_j u_{ij}$$

Wie die gleichnamige Entscheidungsregel oben, so ist auch der Hurwicz-Wert eine lineare Mischung aus dem pessimistischen Wald-Wert und dem absolut optimistischen Wert, der davon ausgeht, daß diejenige Antwort richtig ist, welche, bei adäquater Handlung, den größten Gewinn ergibt. Je größer λ ist, desto optimistischer ist S bezüglich des Ergebnisses.

(3) Da bei der Savage-Niehans Regel die Antworten nach ihrem Bedauern gemessen werden, ist dieses Vorgehen nicht als Grundlage zu nehmen für den Wert einer Frage, der an dem aus der Frage resultierenden Nutzen gemessen werden soll.

Wie oben setzen der Wald- und der Hurwicz-Wert nicht die Existenz subjektiver Wahrscheinlichkeiten voraus. Im Gegensatz dazu steht der

(4) *Bernoulli-Wert*

$$W(f) = \sum_j p_j \max_i u_{ij} = \sum_j p_j u_{jj}$$

Hier wird davon ausgegangen, daß bei voller Beantwortung von f die optimale Entscheidung getroffen wird, so daß sich der Bernoulli-Wert als Erwartungswert der optimalen Ergebnisse darstellen läßt. Ein Spezialfall hiervon ist wie oben der

(5) *Laplace-Wert*

$$W(f) = \frac{1}{|A|} \sum_j \max_i u_{ij} = \frac{1}{|A|} \sum_j u_{jj}$$

als Bernoulli-Wert bei mangelndem Grund.

(6) *Hodges-Lehmann-Wert* mit Vertrauensparameter λ

$$W(f) = \lambda \sum_j p_j \max_i u_{ij} + (1-\lambda) \min_j \max_i u_{ij}$$

Dieser Wert ist die lineare Mischung zwischen Wald- und Bernoulli-Wert, λ spiegelt hierbei das Vertrauen in die gesetzten Wahrscheinlichkeiten wider. Auch hier kann durch mehrfache Mischung ein Wert gefunden werden, der sich den individuellen Bedürfnissen recht gut anpaßt.
Neben dem Wert einer Frage kann mit Hilfe der Nutzenfunktion u nun auch angegeben werden, wieviel S bezahlen muß, um seine Frage an den Informationsträger $T \in \mathcal{T}$ zu stellen. Zuerst schätzt S den Aufwand an Zeit $Z(T)$, Geld $G(T)$ und "Selbstüberwindung" $P(T)$ ab, diese drei Komponenten werden mittels u in eine gewichtete Summe transformiert, die die Kosten angibt, die S aufbringen muß, um die Frage f dem Informationsträger T zu stellen. Dieser Wert wird bezeichnet als *Kosten der Antwort von T auf die Frage f* und abgekürzt geschrieben als $K(f,T)$.

4.2.4 Kostengünstige Fragen

Mit dem Begriff des Wertes einer Frage läßt sich nun auch genauer fassen, wann und ob eine gestellte Frage kostengünstig ist. Grob gesagt ist das Stellen der Frage f an den Informationsträger $T \in \mathcal{T}$ dann kostengünstig, wenn im Verhältnis zum Wert von f soviel aufgewendet wird, wie Informationen im Verhältnis zum Informationsbedarf von f gewonnen werden. Eine exakte Formulierung eben dieses Sachverhaltes erfolgt weiter unten. Zunächst ist noch das Problem zu klären, daß S gar nicht wissen kann, wieviel Informationen T liefern kann, ohne T selber gefragt zu haben und somit die Mittel aufgewandt zu haben, die für eben diese Frage notwendig waren. Dieses Problem wird durch den Begriff der *erwarteten Information* präzisiert.
Beim Befragen eines Informationsträgers $T \in \mathcal{T}$ bekommt S mit Sicherheit eine Antwort, deren Informationsgehalt mit Hilfe der in 2.7, S.32 eingeführten Funktion $I(\alpha, f)$ bestimmt werden kann. Der Wert einer Antwort liegt mit Sicherheit zwischen $-\log n$ und $\log n$, genauer:

$$-(\log n - I(f)) \leq I(\alpha, f) \leq I(f)$$

Für jeden Informationsträger $T \in \mathcal{T}$ hat S eine Plausibilitätsverteilung $w(T)$ auf

$$U = [I(f) - \log n, I(f)] \subset \mathbb{R}$$

Der Erwartungswert von $w(T)$ gibt die Informationsmenge an, mit der S beim Befragen von T rechnet. Überraschungen nach oben und unten sind dabei nicht ausgeschlossen. Im Grenzfall ist $w(T)$ eine Einpunkt-Verteilung, wobei w^* die in diesem Fall sicher erwartete Information angibt, mit der S von T rechnet. Dies liegt z.B. dann vor, wenn S sicher ist, daß T die Frage beantworten kann. Dieser Erwartungswert $E(w(T))$ heißt nun *erwartete Information* und wird mit $E(T)$ abgekürzt.
Von T eine Antwort einzuholen ist für S jetzt preiswert, wenn

$$\frac{E(T)}{I(f)} \geq \frac{K(f,T)}{W(f)}$$

gilt, d.h., wenn im Verhältnis zum eigentlichen Wert der Frage nicht mehr Kosten aufgewendet werden müssen, als Informationen im Verhältnis zum Informationswert der Frage erwartet werden.
Selbstverständlich kann S sich bei seiner Einschätzung von T täuschen und weniger Informationen bekommen als Mittel aufgewandt wurden. Allerdings handelt S zum Zeitpunkt der Frage nach bestem Wissen und Gewissen, wenn obige Abschätzung zugrunde gelegt wird.
Mit Hilfe der Plausibilitätsverteilung $w(T)$ kann auch das (immer subjektive) Risiko abgeschätzt werden, zuviel auszugeben. Dies ist nämlich gleich der Plausibilität, daß

$$\frac{I(\alpha, f)}{I(f)} \leq \frac{W(f)}{K(f, T)}$$

gilt.
Der Einfluß verschiedener Verteilungsfunktionen auf die Auswahl von T hängt somit auch von der Risikobereitschaft von S ab, dieser Aspekt soll hier nicht weiter untersucht werden. Im weiteren beschränke ich mich auf die vereinfachte Annahme, daß die Verteilungsfunktion die oben erwähnte Einpunkt-Verteilung ist.

4.2.5 Entscheidung bei Risiko

Als letztes werde ich den Fall behandeln, der entsteht, wenn es für S keine Möglichkeit mehr gibt, ein Antwort auf seine Frage zu bekommen, d.h. alle Antworten sind entweder zu teuer oder die Menge $\mathcal{T}$ der Informationsträger ist leer. In diesem Fall ist S dennoch gezwungen, sich für eine (wesentliche) Antwort zu entscheiden, so daß der klassische Fall der Entscheidungssituation bei Risiko vorliegt.
Aus jeder Entscheidung von S resultiert ein in Abhängigkeit von dem tatsächlichen Weltzustand, bestimmter Nutzen.
Da jeder Weltzustand mit einer bestimmten Plausibilität eintritt, wird auch der zugehörige Nutzen mit eben dieser Plausibilität erwartet. Jeder Entscheidung entspricht also eine bestimmte Plausibilitätsverteilung des Nutzens. Anstatt, wie oben gefordert, eine bestimmte Anordnung auf dem Antwortenraum A zu finden, können auch die entsprechenden Plausibilitätsverteilungen angeordnet werden. Der Begriff der Antwort wird also durch den der Plausibilitätsverteilung ersetzt (vgl. 2.7, Seite 31). Die hier behandelte Entscheidungssituation besteht dann darin, daß das entscheidende Subjekt S eine Wahl zwischen verschiedenen (diskreten) Plausibilitätsverteilungen treffen muß.
Um möglichst allgemeingültige Entscheidungskriterien zu finden, betrachte ich zunächst die Menge $\mathcal{V}$ aller (diskreten, eindimensionalen) Plausibilitätsverteilungen. Verschiedene Entscheidungen können die gleiche Plausibilitätsverteilung zur Folge haben. Diese Entscheidungen sollen dann als indifferent gelten. Allgemein gilt folgende

Grundannahme: Auf der Menge der Plausibilitätsverteilungen besteht eine Präferenzrelation $\succeq$. Ob eine Plausibilitätsverteilung v_1 einer anderen v_2 vorgezogen wird, hängt einzig und allein von v_1 und v_2 ab und nicht davon, wie diese Plausibilitätsverteilungen zustande gekommen sind.
Diese Grundannahme verlangt nicht die Vergleichbarkeit von je zwei Plausibilitätsverteilungen. Die Unvergleichbarkeit von zwei Verteilungen besagt, daß S nicht weiß, welche Verteilung vorzuziehen ist, nicht weil sie indifferent sind,

sondern weil ihre Konsequenzen nicht vollständig abgesehen werden. Dies widerspricht aber den Vorstellungen eines "homo oeconomicus", der mit minimaler Reaktionszeit minimale Nutzenunterschiede feststellen kann und wird durch folgende Annahme ausgeschlossen:

Ordinales Prinzip: Die Plausibilitätsverteilungen in $\mathcal{V}$ sind durch die Relation $\succeq$ schwach, einfach geordnet, d.h.

1. Aus $v_1 \not\sim v_2$ folgt $v_1 \succ v_2$ oder $v_2 \succ v_1$
2. Aus $v_1 \succeq v_2$ und $v_2 \succeq v_3$ folgt $v_1 \succeq v_3$

Dies ist das gleiche Prinzip wie oben, nur war es dort auf die Antworten bezogen, hier, in Übereinstimmung mit der Grundannahme, auf die Plausibilitätsverteilungen.
Das ordinale Prinzip läßt sich noch schärfer formulieren, mit dem Vorteil einer besseren Handhabung:
Es existiert ein Präferenzfunktional Ψ, das jeder Plausibilitätsverteilung $v \in \mathcal{V}$ eine reelle Zahl zuordnet, so daß für $v_1, v_2 \in \mathcal{V}$ gilt:

$$\Psi[v_1] \geq \Psi[v_2] \Leftrightarrow v_1 \succeq v_2$$

Aus dem ordinalen Prinzip folgt auch eine Präferenzrelation auf dem Nutzen, wenn ein bestimmter Nutzen x als ausgeartete Plausibilitätsverteilung angesehen wird. In diesem Fall ist $x \in \mathcal{V}$. Es kann vorausgesetzt werden, daß ein höherer Nutzen einem niedrigeren vorgezogen wird, dies legt schon der Begriff "Nutzen" nahe.

Normalfall (Monotonieprinzip): Mit $x_1 > x_2$ ist auch $x_1 \succ x_2$.
Dieser Fall liegt auch beim Entscheiden bei Sicherheit vor, der als ausgearteter Risikofall gedeutet werden kann.

Noch ein Prinzip ist zu behandeln, um dem Anspruch rationaler Handlung genügen zu können:

Dominanzprinzip: Sei X_v eine reelle Zufallsvariable mit der Verteilung v und $x' = f(x)$ eine reelle Funktion, die jedem Nutzen x einen günstigeren $x' > x$ zuordnet. v_f sei die Verteilung von $f(X_v)$, die *durch f transformierte Verteilung.* Dann gilt $v_f \succ v$.
v' unterscheidet sich von v nur dadurch, daß mit gleichen Plausibilitäten höherer Nutzen erwartet wird. Damit ist die ganze Risikosituation besser.
Für den oben definierten Normalfall lautet das Dominanzprinzip:

Dominanzprinzip I im Normalfall: Es sei f eine zunehmende Funktion (d.h. $f(x) \geq x$, für alle x) und v_f die durch f transformierte Verteilung von v, dann ist $v_f \succeq v$
Durch Einführen von F_v als Verteilungsfunktion von v kann das Dominanzprinzip noch allgemeiner formuliert werden:

Dominanzprinzip II im Normalfall: Ist $F_{v_1}(x) \geq F_{v_2}(x)$ für alle x, dann ist $v_2 \succeq v_1$.
v_2 ist günstiger als v_1, weil für jedes x die Plausibilität, einen größeren Nutzen als x zu erhalten, größer ist, wenn die Verteilung v_2 zugrunde liegt, als wenn v_1 vorliegt.
Im Dominanzprinzip I werden die Nutzen verbessert, bei gleichbleibenden Plausibilitäten. Im Dominanzprinzip II werden die Nutzen gelassen, dafür die Plausibilitäten so geändert, daß sie für bestimmte höhere Nutzen vergrößert und für niedrigere verkleinert werden. Im ersten Fall liegt Nutzendominanz, im zweiten Plausibilitätsdominanz vor.

4.2.6 Sicherheitsäquivalente

Definition 4.1 *Ein* Sicherheitsäquivalent *einer Plausibilitätsverteilung v ist ein (sicherer) Nutzen $N(v)$, der zu v indifferent ist: $N \sim v$, so daß auch das Präferenzfunktional den gleichen Wert annimmt.: $\Psi[N] = \Psi[v]$.*

So könnte es S egal sein, ob es den Betrag 30 mit Sicherheit bekommt oder ob es die Gelegenheit hat, in einem Spiel mit 40% Plausibilität den Betrag 100 zu gewinnen oder mit 60% nichts zu bekommen. Der Betrag von 30 ist sicher, auf der anderen Seite bietet das Spiel die Möglichkeit, den Betrag von 100 zu bekommen. Es ist durchaus annehmbar, daß die beiden Möglichkeiten von S als indifferent betrachtet werden. In diesem Fall wäre 30 das Sicherheitsäquivalent der Verteilung v mit $v(100) = 0,4$ und $v(0) = 0,6$. $N = 30 \sim v$.
Nicht jede Verteilung besitzt notwendig ein Sicherheitsäquivalent. Ist z.B. S in einer bedrohlichen Lage, aus der es nur durch einen Betrag von 30 entfliehen kann (z.B. durch geeignete Bestechung), so ist jeder kleinere Betrag nutzlos. Zu obiger Verteilung ließe sich kein Sicherheitsäquivalent finden. 30 oder mehr würde S der ungewissen Chance, 100 zu bekommen, vorziehen; jeder kleinere Betrag wäre nutzlos und darum weniger attraktiv als die Möglichkeit, eventuell 100 zu bekommen. Obwohl es in diesem Fall kein Sicherheitsäquivalent gibt, soll diese Annahme durch das folgende Prinzip ausgeschlossen werden:

Stetigkeitsprinzip: Jede Plausibilitätsverteilung besitzt (mindestens) ein Sicherheitsäquivalent.
Obwohl das Stetigkeitsprinzip ausschließt, daß es kein Sicherheitsäquivalent gibt, so bleibt noch die Möglichkeit, daß eine Verteilung zwei Äquivalente hat.

Dieser Fall wird jedoch durch das Monotonieprinzip ausgeschlossen, weil von zwei Nutzen der größere vorgezogen wird. Es gilt also der

Satz 4.1 *Unter Voraussetzung des Monotonie- und des Stetigkeitsprinzips hat jede Plausibilitätsverteilung v genau ein Sicherheitsäquivalent $N(v)$ und dieses kann als Präferenzfunktional aufgefaßt werden.*

□

Damit ist die Suche nach einer Präferenzstruktur umgeändert worden in die Suche nach den entsprechenden Sicherheitsäquivalenten, denn mit der Kenntnis der Sicherheitsäquivalente kann auf die Präferenzstruktur geschlossen werden und umgekehrt.
Das Sicherheitsäquivalent in Situationen, in denen Verlust auftreten kann, ist gleich dem Aufwand, den man auf sich zu nehmen bereit ist, um sich gegen diesen Verlust zu versichern.
Mit Hilfe des Begriffes des Sicherheitsäquivalentes können auch die Grundtypen des Verhaltens bei Risiko charakterisert werden: die *Risikoaversion* und *Risikosympathie*. Erstere bedeutet, daß das Sicherheitsäquivalent kleiner ist als die mathematische Erwartung des Nutzens, letztere, daß es größer als diese ist.

Als eigentlich rationales Entscheidungsprinzip soll nun das Bernoulli-Prinzip vorgestellt werden:
Das Präferenzfunktional von S hat die Gestalt

$$\Psi[v] = E[v(x)] = E_v[x]$$

wobei die Variable x den Nutzen bezeichnet.
Dieses Prinzip basiert im wesentlichen auf der Nutzenfunktion u, die das Ergebnis der Entscheidung für v und eines Zustandes der Welt bewertet.
Das Bernoulli-Prinzip erfüllt das Dominanzprinzip, denn der Übergang von einem Nutzen x zu einem günstigeren x' vergrößert den Erwartungswert und somit das Präferenzfunktional. Im *Normalfall* ist $x = id(x)$ eine streng monoton steigende stetige Funktion. In diesem Fall gilt das *Dominanzprinzip II.*
Nach dem Bernoulli-Prinzip ist das Sicherheitsäquivalent einer Verteilung v gegeben durch

$$N(v) = (E_v[u(y)])$$

wobei y die möglichen Ergebnisse durchläuft, die bei der Verteilung v eintreten können. Da u eine stetige, komponentenweise streng monotone Funktion ist, existiert die Umkehrfunktion u^{-1}. Die Werte von u^{-1} werden i.allg. eine Hyperebene in dem Raum sein, der von den Dimensionen Geld, Zeit und Aufwand aufgespannt wird. Alle Punkte auf dieser Ebene sind allerdings zueinander indifferent, weil ihr Nutzen der gleiche ist. Insofern kann als Sicherheitsäquivalent nicht nur ein zur Verteilung v indifferenter Nutzen angegeben werden, sondern mittels u^{-1} auch eine Menge bezüglich dieses Nutzens indifferenter Ergebnisse.

4.2.7 Axiomatik des Bernoulli-Prinzips

Daß das Bernoulli-Prinzip weitestgehend als rational angesehen wird, verdankt es der Tatsache, aus sehr einsichtigen Axiomen gefolgert werden zu können. Damit entsteht die Möglichkeit, auch in Risikosituationen eine Entscheidung fällen zu können und so rational zu handeln.

1. Das erste Axiom verlangt, daß das ordinale Prinzip gilt, daß also die Präferenzrelation in $\mathcal{V}$ eine schwache, einfache Ordnung induziert.

2. Hier werden einfache Plausibilitätsverteilungen betrachtet, nämlich solche, die mit der Plausibilität p den Nutzen x, mit $1-p$ den Nutzen y versprechen. Diese Verteilung sei mit xpy bezeichnet. Aus $x \succ y$ soll nun folgen, daß $x \succ xpy \succ y$ gilt, sofern $0 < p < 1$, oder sogar, daß $xp_1y \succ xp_2y$, falls $p_1 > p_2$. Diese Forderung stimmt mit dem *Dominanzprinzip II* überein.

3. Dieses Axiom verlangt den stetigen Übergang von x zu y, wenn p von 1 nach 0 abnimmt: Ist $x \succ y \succ z$, dann gibt es ein p, mit $0 < p < 1$, so daß $z \sim xpy$. Diese Bedingung heißt auch *Stetigkeitsaxiom* und ist nicht zu verwechseln mit dem oben angeführten *Stetigkeitsprinzip*.

4. Sind v_1, v_2, v_3 drei Plausibilitätsverteilungen und ist $v_1 \succeq v_2$, dann gilt für die mit einem $p,\ 0 < p < 1$ zusammengesetzten Plausibilitätsverteilungen $v_1pv_3 \succeq v_2pv_3$ und umgekehrt. Diese Bedingung heißt *Substitutionsaxiom* und kann auf folgende Weise veranschaulicht werden:

Links ist ein Glücksspiel dargestellt, daß dem beteiligten Subjekt S in einer ersten Stufe mit der Plausibilität $1-p$ die Verteilung v_3 zukommen läßt, durch die in einer zweiten Stufe der tatsächliche Gewinn oder Verlust ermittelt wird. Mit der Plausibilität p kann S aber auch vor die Wahl gestellt werden, in einer zweiten Stufe zwischen den Verteilungen v_1 und v_2 zu wählen. Im Fall $v_1 \succ v_2$ wird S sich für v_1 entscheiden.

Auf der rechten Seite muß S sich direkt zwischen den beiden Spielen $v_1 p v_3$ und $v_2 p v_3$ entscheiden. Diese Wahl stellt sich aber als gleichwertig zum linken Problem dar. Wenn dort S die Verteilung v_1 der Verteilung v_2 vorgezogen hat, so wird S hier das Spiel $v_1 p v_3$ dem Spiel $v_2 p v_3$ vorziehen, unabhängig von der Verteilung v_3 und unabhängig von der Größe p, wobei $p > 0$ angenommen wird. Das gleiche gilt im Fall $v_1 \sim v_2$. Anstatt sich erst nach der ersten Stufe zu entscheiden, kann S direkt die Wahl zwischen verschieden Spielen treffen. Die Tatsache, welches Spiel gespielt wird, ist für S nicht entscheidend. Dieses Axiom kann somit auch als fehlende Lust, dem Spieltrieb nachzugehen, interpretiert werden und ist somit kennzeichnend für rationales Handeln in Ungewißheitssituationen.
Hierbei ist Axiom 2 in Axiom 4 enthalten.

Diese Axiome reichen nun hin, um das eingeführte Bernoulli-Prinzip zu bestätigen. Der Beweis wird hier nur für beschränkte, diskrete Verteilungen gezeigt. Diese Verteilungen sind gerade die für diese Arbeit interessanten.
Zunächst muß die Nutzenmatrix $\mathcal{U}$ geeignet transformiert werden: Seien $\underline{u}$ der minimale und $\overline{u}$ der maximale Wert der Nutzenmatrix. Jeder Eintrag u_{ij} der Nutzenmatrix wird nun transformiert zu

$$\hat{u}_{ij} := \frac{u_{ij} - \underline{u}}{\overline{u} - \underline{u}}$$

Die Einträge der so transformierten Nutzenmatrix liegen alle zwischen 0 und 1, $0 \leq \hat{u}_{ij} \leq 1$, eine Präferenzstruktur auf den Zeilen von $\mathcal{U}$ bleibt bei dieser Transformation erhalten. Für jede Verteilung v liegen die Nutzen $u_1, u_2, \ldots, u_n$ zwischen $\underline{u}$ und $\overline{u}$, d.h. die transformierten Nutzen $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n$ zwischen 0 und 1. Die Plausibilitäten dieser Nutzen seien mit $p_1, p_2, \ldots, p_n, \sum p_i = 1$ bezeichnet. Es gilt $\hat{u}_i \sim 1\hat{u}_i 0$. Mit dem Substitutionsaxiom kann in der Verteilung für $\hat{u}_i$ die indifferente Verteilung $1\hat{u}_i 0$ eingesetzt werden. Hieraus entsteht eine zusammengesetzte Verteilung, die in der ersten Stufe mit der Plausibilität p_i die Verteilung $1\hat{u}_i 0$, in der zweiten Stufe mit der Plausibilität $\hat{u}_i$ den Nutzen 1 eintreten läßt. Der Nutzen 1 tritt also mit der Plausibilität

$$\overline{v} := \sum_{i=1}^{n} p_i u_i$$

ein, der Nutzen 0 mit der Gegenplausibilität. Dieser Term ist gerade der Erwartungswert des Nutzens von v. Die zusammengesetzte und damit auch die ursprüngliche Verteilung ist somit indifferent zu $1(\sum p_i u_i)0$. Um zwei Verteilungen v_1 und v_2 zu vergleichen, reicht es also, die zu ihnen jeweils indifferenten Verteilungen $1\overline{v_1}0$ und $1\overline{v_2}0$ zu vergleichen. Nach dem zweiten Axiom gilt aber $1\overline{v_1}0 \succeq 1\overline{v_2}0$ genau dann, wenn $\overline{v_1} \geq \overline{v_2}$ gilt. Folglich werden die Verteilungen nach ihren Erwartungswerten geordnet. Genau dies ist die Aussage des Bernoulli-Prinzips.

Die so erhaltene Präferenzstruktur, gewonnen mit Hilfe der Matrix $\hat{\mathcal{U}}$, läßt sich ohne Schwierigkeiten auf die ursprüngliche Nutzenmatrix $\mathcal{U}$ übertragen, weil die Plausibilitäten sich bei der Transformation nicht geändert haben, lediglich die Nutzenwerte.
Damit ist eine Entscheidungsregel gefunden, die es ermöglicht, sich im Risikofalle, d.h. im Falle unvollständiger Antwort, für eine Antwort zu entscheiden. Zusätzlich kann bei einer solchen Entscheidung auch noch das *Risiko* ρ gemessen werden, als Quotient von tatsächlicher Desinformation und maximaler Desinformation.

$$\rho = \frac{H(f)}{\log_n(n)}, \text{ mit } n = grad(f)$$

Für diesen Wert gilt:

$$0 \leq \rho \leq 1$$

bzw. bei einer Multiplikation mit 100 kann das Risiko auch in Prozent angegeben werde. Es ist 0, wenn die Frage vollständig beantwortet ist und S bei Sicherheit handelt. Natürlich kann S sich getäuscht haben oder eine Fehlinformation und somit eine falsche Antwort bekommen haben. Dennoch ist auch in diesem Fall das Gefühl der Sicherheit in S vorhanden. Diese Sicherheit σ läßt sich als fehlendes Risiko interpretieren:

$$\sigma = 1 - \rho$$

S handelt also in Unsicherheit, wenn $\sigma < 1$ und nur im Falle $\sigma = 1$ im Zustande der Sicherheit. Ziel des ganzen Entscheidungsproblems ist es also, σ möglichst groß und damit ρ möglichst klein werden zu lassen.
S kann sich also auch entscheiden, die zur Verfügung stehenden Kosten dazu einzusetzen, das Risiko unter einen vorgegebenen Wert zu bringen, selbst wenn die volle Antwort nicht möglich ist. Eine solche Antwort kann dann immer noch als kostengünstig angesehen werden.
Das Bernoulli-Prinzip kann als Grundlage für rationales Handeln angesehen werden, was nicht besagt, daß es jeder Handlung zugrundeliegt. Ähnlich wie die Gesetze der Logik als Lehre vom rationalen Schließen nicht bei jeder Schlußfolgerung beachtet werden müssen, so muß nicht jede Entscheidung nach dem Erwartungswert ihres Nutzens ausgewählt werden. Allerdings werden Handlungen, die auf nicht-logischen Schlüssen basieren, als unrational angesehen. Ebenso können Entscheidungen, die nicht nach dem Bernoulli-Prinzip gefällt werden, als unrational eingestuft werden. Ich möchte die Diskussion über die Rationalität des Bernoulli-Prinzips an dieser Stelle nicht weiterführen, sondern vielmehr auf eine in Kapitel 3 eingeführte Anwendung zurückkommmen:

In 3.5, S.49 wurde bereits das Problem angesprochen, eine Entscheidungsregel zu finden, die eine minimale Fehlerplausibilität beim Zuordnen von Ausgabe- zu

Eingabezeichen garantiert. Die Nutzenmatrix für dieses Problem kann mit Eingabealphabet $A = \{a_1, a_2, \ldots, a_n\}$ folgendermaßen aussehen. Wenn S annimmt, daß Zeichen a_i gesendet wurde und tatsächlich a_i gesendet wurde, so trägt dies den Nutzen 1. In allen anderen Fällen ist der Nutzen 0. Die Nutzenmatrix sieht also folgendermaßen aus:

	p_1	p_2	$\cdots$	p_n
	a_1	a_2	$\cdots$	a_n
a_1	1	0	$\cdots$	0
a_2	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_n	0	0	$\cdots$	1

Die Entscheidungsregel 3.5 auf Seite 51 besagte, sich für das Zeichen zu entscheiden, welches am plausibelsten ist. Hierfür ist die Fehlerplausibilität nach Satz 3.8 minimal. Diese Regel ist identisch mit der Bernoulli-Regel, welche besagt, sich für die Antwort zu entscheiden, die den größten Nutzen erwarten läßt. Der Erwartungswert der Antworten a_i ist in der obigen Nutzenmatrix die Plausibilität für das Auftreten von Zeichen a_i. Die Antwort mit dem größten Nutzen-Erwartungswert ist also die Antwort, die am plausibelsten ist, was i.allg. allerdings nicht der Fall ist (4.2.2, S.60)

5 Abschließende Bemerkungen

In der vorliegenden Arbeit wurde gezeigt, daß sich viele Aspekte rationalen Verhaltens aus einsichtigen Axiomen herleiten lassen. Angefangen von den Axiomen der Plausibilitätstheorie, interpretiert als Einschätzung unter Androhung von Sanktionen, über das Informationsmaß, logisch richtigen Schließens, bis hin zu rationalem Verhalten in Risikosituationen.
Um das Ziel, Risikosituationen durch Fragen zu verbessern, zu erreichen, habe ich mich lediglich auf die im folgenden noch einmal aufgelisteten Axiome gestützt. Insofern sie als vernünftig und rational eingestuft werden, läßt sich das gesamte Vorgehen als rational ansehen.

Die Axiome sicherer Logik aus 4.1.1 als Bedingungen rationalen Schließens:

1. $A = A$ (Satz von der Identität)
2. $A = \neg\neg A$ (Satz vom Widerspruch)
3. Jedes x ist entweder A oder $\neg A$ (Satz vom ausgeschlossenen Dritten)

Die Axiome unsicherer Logik aus 1.7 als Bedingungen kohärenter Einschätzung:

1. nicht-Negativität: Gilt sicher $X \geq 0$, so ist sicher $p(X) \geq 0$
2. endliche Additivität der Plausibilitätsfunktion p:

$$p(X + Y) = p(X) + p(Y)$$

3. Für bedingte Plausibilitäten gilt folgende Ungleichung:

$$p(E|H) = p(EH)/p(H), \;\; p(X|H) = p(XH)/p(H)$$

Die Axiome zur Bestimmung des Informationsgehaltes $H(p_1, p_2, \ldots, p_n)$ einer Sachinformationfrage aus 2.4

1. $H(p_1, \ldots, p_n)$ ist stetig
2. $$H(p_1, \ldots, p_n) \leq H(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n-mal}).$$
3. $$\begin{aligned} H(p_1, \ldots, p_n) &= H(p_1, \ldots, p_{i-1}, p_i + p_{i+1}, p_{i+2}, \ldots, p_n) \\ &+ (p_i + p_{i+1}) H\left(\frac{p_i}{p_i + p_{i+1}}, \frac{p_{i+1}}{p_i + p_{i+1}}\right). \end{aligned}$$

Die Axiome für eine rationale Entscheidung für eine Nutzenverteilung in Risikosituationen aus 4.2.7:

1. Die Präferenzrelation $\succeq$ induziert in der Menge $\mathcal{V}$ aller endlichen, diskreten Verteilungen eine schwache, einfache Ordnung (ordinales Prinzip).
2. ist $x \succ y \succ z$, dann gibt es ein p mit $0 < p < 1$, so daß $z \sim xpy$ gilt (Dominanzprinzip II).
3. Sind v_1, v_2, v_3 drei Plausibilitätsverteilungen und ist $v_1 \succeq v_2$, dann gilt für die mit einem $p, 0 < p < 1$ zusammengesetzten Plausibilitätsverteilungen $v_1 p v_3 \succeq v_2 p v_3$ und umgekehrt (Substitutionsaxiom).

Was ist dadurch gewonnen? Sicherlich nicht, daß ein Subjekt sich in einer Entscheidungssituation hinsetzt und ausrechnet, welches die beste Entscheidung wäre. Das Kalkül wird wohl nur in besonderen Fällen explizit zum Einsatz kommen. Nicht selten kostet das Aufstellen der Entscheidungsmatrix mehr Aufwand, als die Frage wert ist.
Mit den hier eingeführten Begrifflichkeiten lassen sich Entscheidungssituationen beschreiben, im Gegensatz zum diffusen Gefühl der Unsicherheit. Die Frage

„Was soll ich tun ?“ ist umgeformt in die Fragen „Was ist mir das Ganze eigentlich wert ?“, „Wie groß ist das Risiko ?“ und „Was habe ich davon, so oder so zu handeln ?“. Die Unsicherheit wird begrifflich gefaßt und somit greifbarer, beschreibbarer. Das Weltganze ist analytisch zerlegt und wird handlicher. Das ist das eigentliche Ziel der hier vorgestellten Modellierung. Der Algorithmus aus Kapitel 4 soll nicht die eigene Urteilsfähigkeit ersetzen, sondern begriffliche Klarheit schaffen.
Modellierung ist immer auch Reduktion. Der wichtigste Aspekt des vorliegenden Modells ist Rationalität, kondensiert in den oben angeführten Axiomen. Ob diese Axiome als Handlungsgrundlage genommen werden, oder ob sie lediglich als Vorschlag interpretiert werden, liegt als Entscheidung weiterhin beim Entscheidungsträger. Die Entscheidung für oder gegen eine Handlungsweise setzt aber die Kenntnis dieser Weise voraus. Das hier behandelte Modell eröffnet die Möglichkeit rationaler Entscheidungen nach strengen Kriterien oder eben auch ihre Nicht-Einhaltung. Insofern will diese Arbeit dazu beitragen, die Freiheit in der Entscheidung dem Subjekt zurückzugeben und nicht, sie ihm abzuhnemen.

Literatur

[Ba] Bamberg, Günter. *Statistische Entscheidungstheorie,* Würzburg: Physica 1972

[De] Denis-Papin, Maurice; Cullmann, Georges. *Übungsaufgaben zur Informationstheorie,* Braunschweig: Vieweg 1972

[Df] Definetti, Bruno. *Theorie of Probability,* New York: Wiley 1974

[Di] Dinkelbach, Werner. *Entscheidungsmodelle,* Berlin; New York: de Gruyter 1982

[Fl] Flechtner, Hans-Joachim. *Grundbegriffe der Kybernetik,* München: dtv 1984

[Fr] Freytag-Löringhoff, Bruno Baron von. *Logik,* Stuttgart; Köln: Kohlhammer 1955

[He] Henze, E.; Homuth, H.H. *Einführung in die Informationstheorie,* Braunschweig: Vieweg 1970^{3}

[Ja] Jaglom, A.M.; Jaglom I.M. *Wahrscheinlichkeit und Information,* Berlin: VEB 1967

[Ka] Kant, Immanuel. *Kritik der reinen Vernunft,* Frankfurt/Main: Suhrkamp 1974^{12}

[Kl] Klausmann, Hans Siegfried. *Stochastische Entscheidungsbäume,* Meisenheim am Glan: Hain 1976

[PM] Peter Moosleitner, Gerhard. *Logik-Trainer 2,* München: Gruner+Jahr 1988

[Ra] Ramsey, Frank Plumpton. *The Foundations of Mathematics,* London: Routledge&Kegan Paul 1931^{4}

[Sa] Savage, Leonard J. *The Foundations of Statistics,* New York: Wiley 1954

[Sc] Schneeweiß, Hans. *Entscheidungskriterien bei Risiko,* Berlin; Heidelberg: Springer 1966

[Se] Seiffert, Helmut; Radnitzky, Gerard (Hrsg.). *Handlexikon zur Wissenschaftstheorie,* München: dtv 1992

[Sh] Shannon, Claude; Weaver, Warren. *Mathematische Grundlagen der Informationstheorie,* München; Wien: Oldenbourg 1976

[Tz] Tzschach, Hans; Haßlinger, Gerhard. *Codes für den störungssicheren Datenverkehr,* München; Wien: Oldenbourg 1993

[Ur] Ursul, A.D. *Information,* Berlin(Ost): Dietz 1970

[Wi] Wittgenstein, Ludwig. *Tractatus logico-philosophicus,* Frankfurt/Main: Suhrkamp 1990[7]