Moritz Müller[1]

# Ein Vergleich zweier ästhetischer Prinzipien

## *1. Einleitung.*

Edgar Allan Poes Detektivgeschichte „The Murders in the Rue Morgue“ besteht aus zwei Teilen. Der erste Teil beschreibt den Kriminalfall und besteht im Wesentlichen in der Angabe einer Liste von Daten, wer, wann, wo und wie getötet worden ist. Das ist alles sehr schrecklich und verwirrend und wird im zweiten Teil einer rationalen Analyse unterworfen, die schließlich zur Klärung des Falles führt. Was ist schön an solch einer Geschichte?

Poes Geschichte ist nicht dadurch eine schöne, dass sie etwas schön darstellt. Auch informiert sie den Leser nicht in bestimmter Weise über einen Sachverhalt, wie Kriminalgeschichten, die unter anderem den Zweck haben, etwas darzustellen, etwa eine soziale Schicht, eine Stadt oder eine Persönlichkeit. Poes Geschichte gefällt auch nicht in der Weise, in der eine clevere Metapher gefällt, sie ist keineswegs symbolisch, sie macht keine, sagen wir, politische oder philosophische Aussage. Sie hängt gleichsam in der Luft, sie ist weltfremd, sie besteht *nur* in der Angabe eines Falls, also einer Frage, und dessen Klärung, also deren Antwort. Verglichen mit Kommissar Brunettis 100. Fall, erscheint Poes Geschichte wie eine Denksportaufgabe oder eine „mathematische Textaufgabe“ in einem Grundschulbuch – inklusive Lösung. Schön kann daran nur die Klärung als solche sein.

In welchem Sinne können Erklärungen schön sein? Nach den Erläuterungen von Magnus Klaue[2] gibt es zwei ästhetische Prinzipien der frühen englischen Kriminalliteratur, zwei Prinzipien, die jeweils ein Kriterium für die Schönheit von Erklärungen liefern, und die hier folgendermaßen wiedergeben werden:

**Prinzip 1**
*Eine Erklärung ist umso schöner, je wahrscheinlicher sie ist.*

**Prinzip 2**
*Eine Erklärung ist umso schöner, je einfacher sie ist.*

Angenommen wir geben eine Erklärung des Falls Rue Morgue à la Däniken durch irgendeine langwierige Geschichte über eine Invasion Außerirdischer. Vermutlich hätte Poe dieser Vorschlag nicht gefallen, jedenfalls wäre eine solche Erklärung gemäß keinem der beiden obigen Prinzipen besonders schön.

Poes Erklärung dagegen ist gemäß beiden Prinzipien eine recht schöne. Jedoch ergeben sich bei genauerer Betrachtung der beiden Prinzipien zumindest drei Probleme: Man mag etwa einwenden, dass es, intuitiv gesehen, nicht gerade wahrscheinlich ist, von einem wildgewordenen Affen in Stücke gerissen zu werden. Der Begriff der Wahrscheinlichkeit muss hier aber als Wahrscheinlichkeit in Bezug auf den gegeben Fall verstanden werden. Außerdem haben wir es, wie gesagt, mit einer ‚in der Luft hängenden’, nicht unbedingt realistischen Geschichte zu tun, so dass der Sinn, in dem hier etwas wahrscheinlich ist oder nicht, wohl kaum der intuitive, empirische, durch den ‚common sense‘ alltäglicher Erfahrung gegebene Sinn ist. Der hier verwendete Wahrscheinlichkeitsbegriff muss in irgendeinem Sinne apriorisch sein. In der Frage, was denn eine apriorische Wahrscheinlichkeit sein soll, liegt die erste Schwierigkeit.

---

1 Centre de Recerca Matemàtica (CRM), Barcelona; der Verfasser dankt der John Templeton Foundation für die Unterstützung durch das CRM Infinity Project, Grant #13152: „The Myriad Aspects of Infinity“.
2 Magnus Klaue war Tagungsteilnehmer, ihm danke ich für viele geduldige Erklärungen. Ebenso danke ich Markus Junker für detaillierte Kommentare zu einer früheren Version dieser Arbeit.

Poes Erklärung ist offenbar eine recht einfache. Es ist aber zweitens wiederum gar nicht so leicht zu erklären, in welchem Sinne Poes Erklärung einfacher ist als etwa die à la Däniken. Es ist im Allgemeinen nicht klar, nach welchen Kriterien eine Erklärung als einfacher als eine andere gelten soll.
Eine dritte Schwierigkeit, die genannten beiden Prinzipien zu verstehen, besteht darin, zu erklären, was man mit ‚Erklärung' meint. Aber nehmen wir einmal an, wir hätten alle diese Schwierigkeiten überwunden und würden die beiden Prinzipien mit zuversichtlicher Sicherheit verstehen und anwenden können. Wie verhalten sich die beiden Prinzipien zueinander? Intuitiv scheint es sowohl Erklärungen zu geben, die gemäß keinem der beiden Prinzipien schön sind, so wie etwa obige Erklärung à la Däniken, und es scheint Erklärungen zu geben, die gemäß beiden Prinzipien schön sind, wie etwa Poes Erklärung. Gibt es Erklärungen, die gemäß dem ersten, nicht aber gemäß dem zweiten Prinzip schön sind? Oder umgekehrt?

Wir möchten im Folgenden eine Lesart der beiden Prinzipien vorschlagen und argumentieren, dass sie äquivalent sind. Wir werden dabei folgendermaßen vorgehen: In der oben gegebenen Formulierung lassen sich die beiden ästhetischen Prinzipien mit Begriffen aus der algorithmischen Informationstheorie formalisieren. Dies ist nicht neu, sondern eine Übertragung von Argumenten aus Arbeiten von Min Li und Paul M. B. Vitányi.[3] In Abschnitt 2 geben wir eine Formalisierung davon, was es heißt, eine möglichst ‚wahrscheinliche' Erklärung zu wählen. In Abschnitt 3 geben wir eine Formalisierung davon, was es heißt, eine möglichst ‚einfache' Erklärung zu wählen. Zu erklären, was eine Erklärung ist, ist ein ebenso bekanntes wie schwieriges Problem. Deswegen werden wir nicht versuchen, das zu tun. Stattdessen ziehen wir in Abschnitt 4 verschiedene Schlussfolgerungen, je nachdem, was man unter einer ‚Erklärung' verstehen will. Schließlich machen wir in Abschnitt 5 noch eine methodische Bemerkung.

## *2. Über das erste Prinzip.*

### 2.1. Bayesianismus.

Das Induktionsproblem (Sextus Empiricus, David Hume) besteht darin, wie man, ausgehend von gegebenen Beobachtungen oder Daten, eine Hypothese wählen soll, um die gegebenen Daten zu erklären. Ein naheliegender Vorschlag, den wir im Folgenden den *bayesianischen* Vorschlag nennen wollen, ist der, eine Hypothese zu wählen, die gegeben die Daten eine möglichst hohe Wahrscheinlichkeit hat. Dies ist gerade das, was Prinzip 1 sagt.

Um dies formaler auszudrücken, nehmen wir an, dass alle möglichen Hypothesen H und alle möglichen Daten D eine gewisse Wahrscheinlichkeit Pr[H] bzw. Pr[D] haben. In anderen Worten, wir nehmen an, dass sowohl Hypothesen als auch Daten Ereignisse in einem geeigneten Wahrscheinlichkeitsraum sind, und bezeichnen das Wahrscheinlichkeitsmaß mit ‚Pr'. Die

3 Paul M.B. Vitanyi, Min Li: Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity, in: IEEE Transactions on Information Theory, 46, 2000, 2, S. 446–464; dies.: Ideal MDL and Its Relation to Bayesianism, in: Proceedings ISIS: Information, Statistics and Induction in Science World Scientific, Singapur 1996, S. 282–291; dies.: Inductive Reasoning, in: DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 17, 1994 (Language Computations; Proceedings DIMACS Workshop on Human Language, 20.–22. März 1992), S. 127–148. Einführungen in die algorithmische Informationstheorie geben dies.: An Introduction to Kolmogorov Complexity and Its Applications, 3. Aufl., New York 2008, und Cristian S. Calude: Information and Randomness. An Algorithmic Perspective, 2. Aufl., Berlin, Heidelberg 2002.

Wahrscheinlichkeit $Pr[E_1|E_2]$ eines Ereignisses $E_1$ gegeben ein Ereignis $E_2$ ist definiert durch den Bruch $Pr[E_1$ und $E_2] / Pr[E_2]$ (wobei man annehmen muss, dass $E_2$ nicht Wahrscheinlichkeit 0 hat).
Die Bayesianer bitten uns, gegeben die Daten D, eine Hypothese H so zu wählen, dass Pr[H|D] möglichst groß ist. Um diese Aufforderung besser zu verstehen, schreiben wir etwas um und erhalten die sogenannte *bayessche Regel*[4]:

(1) Pr[H|D]
= (Pr[H und D] / Pr[D]) · (Pr[H] / Pr[H])
= (Pr[H und D] / Pr[H]) · (Pr[H] / Pr[D])
= Pr[D|H] · Pr[H] / Pr[D].

Um den Vorschlag der Bayesianer zu verstehen, versuchen wir die Terme, die hier auf der rechten Seite der bayesschen Regel (1) vorkommen, zu interpretieren. Wir haben die Daten D beobachtet und sollen jetzt eine bestimmte Hypothese H wählen – wir können also den konstanten Faktor Pr[D] der von H gar nicht abhängt, getrost vergessen. Bleiben die Faktoren Pr[D|H] und Pr[H]. Der Faktor Pr[D|H] ist die Wahrscheinlichkeit, unsere Daten zu beobachten, unter der Annahme, die Hypothese treffe zu. Nennen wir diesen Wert den *Erklärungswert* der Hypothese H für die Daten D. Wir diskutieren diesen Faktor später.
Der Faktor Pr[H] ist die Wahrscheinlichkeit der Hypothese unabhängig von irgendwelchen Beobachtungen: die sogenannte *a priori Wahrscheinlichkeit* der Hypothese H. Der Bayesianische Vorschlag besagt, wir mögen zur Erklärung gegebener Daten eine a priori möglichst wahrscheinliche Hypothese mit möglichst hohem Erklärungswert wählen. Das klingt zwar vernünftig, aber was sind denn diese a priori Wahrscheinlichkeiten? Darüber mag von Fall zu Fall Streit ausbrechen. Gegner des Bayesianismus werfen diesem sogar einen circulus vitiosus vor: Um mittels dem Bayesianismus ein Phänomen zu verstehen, muss man a priori Wahrscheinlichkeiten kennen; aber damit wird bereits ein Wissen über das Phänomen vorausgesetzt, das zu verstehen doch gerade das Ziel des ganzen Unterfangens ist.

Ein Ausweg aus dieser Krise wäre den Bayesianern dadurch gegeben, wenn man in vernünftiger Art und Weise die a priori Wahrscheinlichkeiten ein für alle mal festlegen könnte, ganz allgemein und unabhängig vom konkreten Fall. Das wäre dann auch ein Wahrscheinlichkeitsbegriff, der in einem geeigneten Sinne apriorisch ist, so wie dies in Prinzip 1 in der Einleitung gefordert wird.
Ein solcher Vorschlag scheint utopisch, wurde aber von Solomonoff vorgelegt. Wir erklären Solomonoffs Vorschlag und verweisen auf die Literatur[5] zu Ergebnissen dazu, in welchem Sinn sein Vorschlag in der Tat ein vernünftiger ist:

> Essentially combining the ideas of Ockham, Bayes, and modern computability theory, R.J. Solomonoff has successfully invented a ‚perfect‘ induction theory. First, combine Occam's razor principle and modern computability theory to obtain Kolmogorow complexity. With Kolmogorow complexity define a *universal prior* which dominates, up to a multiplicative constant, all computable prior probability distributions. Use this universal prior in Bayes' Rule substituting it for *any* computable prior probability which may actually hold. This results in a general theory of inductive inference.[6]

4 Wir nehmen hier an, dass Pr[H] > 0 – Hypothesen, die sicherlich (mit Wahrscheinlichkeit 1) nicht zutreffen, sind uns gleichgültig. Außerdem nehmen wir an, dass Pr[D] > 0, also dass wir nicht gerade Daten beobachtet haben, die Wahrscheinlichkeit 0 haben.
5 Vitanyi, Li: Description (Anm. 4), dies.: Reasoning (Anm. 4).
6 Vitanyi, Li: Reasoning (Anm. 4), S. 3.

## 2.2. Solomonoffs universales a priori.

Wir wechseln zu einer etwas abstrakteren Sicht der Dinge. Was sind denn Daten und Hypothesen? Nun, was auch immer sie sind, sie lassen sich in irgendeiner Form sprachlich beschreiben. Eine solche Beschreibung ist eine (endliche) Folge von Zeichen in einem (endlichen) Alphabet. Weil es für unsere Zwecke gleichgültig ist, nehmen wir der Einfachheit halber an, das Alphabet bestehe aus nur zwei Buchstaben, 0 und 1. Eine endliche Folge von Nullen und Einsen ist ein *binärer String*. Ein Wahrscheinlichkeitsmaß Pr ordnet jedem binären String $x$ seine Wahrscheinlichkeit Pr[$x$] zu, eine reelle Zahl zwischen 0 und 1. Für die bedingten Wahrscheinlichkeiten, die in der bayesschen Regel (1) vorkommen, erklären wir für jeden binären String $y$ das Wahrscheinlichkeitsmaß Pr[ · |$y$] durch die Festsetzung Pr[$x$|$y$] = Pr[$<x,y>$]. Hierbei ist $<x,y>$ ein geeigneter String, der das geordnete Paar ($x$,$y$) der Strings $x$ und $y$ repräsentiert.
Wir machen hier zwei technische Annahmen: Normalerweise wird verlangt, dass sich die Wahrscheinlichkeiten zu 1 summieren, das heißt, dass $\sum_x \Pr[x] = 1$; wir verlangen hier aber nur, dass $\sum_x \Pr[x] \le 1$. Außerdem verlangen wir, dass die Zuordnung Pr in geeignetem Sinne effektiv ist.[7] Solche Zuordnungen Pr nennen wir *effektive Wahrscheinlichkeitsmaße*. Jetzt kann man folgendes bemerkenswertes Theorem beweisen:

**Theorem 1**
*Es gibt ein* universales *effektives Wahrscheinlichkeitsmaß* **m**, *das heißt, für jedes effektive Wahrscheinlichkeitsmaß* Pr *gilt* $\Pr[x] \le O(\mathbf{m}[x])$.

Hierbei bedeutet $\Pr[x] \le O(\mathbf{m}[x])$, dass es eine positive Konstante *const* gibt (die nur vom Maß Pr abhängt) so dass $\Pr[x] \le const \cdot \mathbf{m}[x]$ für alle binären Strings $x$.

Solomonoffs Vorschlag besteht darin, den bayesianischen Vorschlag unter Verwendung von **m** als a priori Wahrscheinlichkeit zu verwenden. Das läuft darauf hinaus, eine Hypothese H zur Erklärung gegebener Daten D so zu wählen, dass Pr[D|H] · **m**[H] möglichst groß ist. Unsere Formalisierung von Prinzip 1 lautet also:

**Formalisierung von Prinzip 1**
*Eine Erklärung* H *für Daten* D *ist umso schöner, je größer* Pr[D|H] · **m**[H] *ist.*

# *3. Über das zweite Prinzip.*

## 3.1. Ockhams Rasiermesser.

Warum sagen wir unseren enttäuschten Kindern, dass es keine Einhörner gebe? Woher wissen wir das? Die Behauptung, es gäbe sie, ist verträglich mit all unserem sonstigen Wissen. Dass es Einhörner gibt, mag intuitiv zwar als unwahrscheinlich gelten, da ja bisher keine Hinweise auf ihre Existenz gefunden wurden, doch unmöglich ist es nicht. Trotzdem ist es natürlich unsinnig zu behaupten, Einörner existierten – aber in welchem Sinn ist das unsinnig? Ein Grund gegen die Existenz von Einhörnern ist, dass es keinen Grund für sie gibt. Anders gesagt, alle Daten können, wenn überhaupt, genauso gut ohne Einhörner erklärt werden. In diesem Sinne ist die Annahme, es gäbe Einhörner, überflüssig und deswegen abzulehnen. Das scheint ein guter Grund gegen die Existenz von Einhörnern zu sein und vielleicht gibt es auch keinen besseren.

Wir haben die Theorie mit und die Theorie ohne Einhörner gemäß einem Kriterium verglichen, das

7 Wir verlangen Folgendes: Es gibt einen Algorithmus, der gegeben einen binären String $x$ und (eine Kodierung einer ) eine natürliche(n) Zahl $n$ eine (Kodierung einer) rationale(n) Zahl K($x$,$n$) ausgibt, so dass für alle $x$ die Folge K($x$,$n$ ), $n \ge 0$, schwach monoton wächst (also $K(x,0) \le K(x,1) \le K(x,2) \le \ldots$) und gegen Pr[$x$] konvergiert.

als ‚Ockhams Rasiermesser' bekannt ist: *Entia non sunt multiplicanda sine necessitate.* Das Kriterium wird oft so ausgelegt, dass man von zwei Theorien (die sich, sagen wir, gegenüber den empirischen Daten gleich verhalten) die ‚einfachere' bevorzugen solle. In welchem Sinne eine Theorie ‚einfacher' als eine andere ist, mag zwar wie in dem obigen, so in vielen Fällen auf der Hand liegen, im Allgemeinen ist das aber schwierig zu erklären. Der algorithmischen Informationstheorie ist es jedoch gelungen, den Begriff der Einfachheit durch den Begriff der Kolmogorow-Komplexität zu formalisieren.

## 3.2. Kolmogorow-Komplexität.

Wir haben uns bereits darauf geeinigt, unter Hypothesen binäre Strings zu verstehen. Gesucht ist also ein Konzept der Einfachheit für binäre Strings, wie etwa die Folgenden:

000000000000000000000000000000000000000000000000000000
000010101010101010101010101010101010101010101010101010
010101101111001000001111101010011100101010010100111100

Intuitiv ist der erste String besonders einfach: 54 mal 0, und der zweite ebenfalls: erst 4 mal 0, dann 25 mal 10. Der dritte String ist in dem Sinne komplizierter als die anderen, dass er keine so kurze Beschreibung hat. Intuitiv kann eine Beschreibung umso kürzer ausfallen, je mehr Regelmäßigkeiten es im String gibt, also je weniger ‚zufällig' er ist.
Wir nehmen diese Intuition als Ansatz zu einer Formalisierung des Einfachheitsbegriffes. Eine „Beschreibung" eines binären Strings $x$ ist dabei ein „Code" von $x$ oder allgemeiner ausgedrückt, eine „Vorschrift", ein „Algorithmus", ein „Programm", die/der/das $x$ generiert.

Wir definieren wie folgt: die *Kolmogorow-Komplexität* $\mathrm{K}(x)$ von $x$ ist die Länge eines kürzesten Programms für $x$. Dann ist ein binärer String umso einfacher, je kleiner seine Kolmogorow Komplexität ist.
Unter einem Computer verstehen wir wie üblich formal eine Turingmaschine. Ein Programm für $x$ ist per definitionem ein binärer String $p$, so dass ein fest gewählter universaler präfixfreier Computer bei Input $p$ mit Output $x$ hält. Universalität heißt nur, dass die Maschine alle möglichen Programme ausführen kann, also dass man auf ihr jede mögliche Software implementieren kann. Die Annahme der Präfixfreiheit ist eine technische und braucht uns hier nicht zu beunruhigen; grob gesagt, erlaubt sie nur Programme, die ihr eigenes Ende angeben, etwa indem sie mit einer Halteanweisung oder Klammer-Schließung enden[8]. Die Definition von $\mathrm{K}(x)$ hängt nicht wirklich von der Wahl des universellen Computers ab: Für je zwei solcher Computer unterscheiden sich die zugehörigen Werte $\mathrm{K}(x)$ nur um höchstens eine Konstante.
Wir benötigen noch die bedingte KolmogorowKomplexität: Für einen String $y$ ordnet $\mathrm{K}(\,\cdot\,|y)$ jedem String $x$ seine Kolmogorow-Komplexität $\mathrm{K}(x|y) = \mathrm{K}(<x,y>)$ *bezüglich* $y$ zu (hier verwenden wir wieder die Kodierung $<\cdot\,,\cdot>$ von Paaren aus Abschnitt 2.2).

Wir erreichen die

**Erste Formalisierung von Prinzip 2**
*Eine Erklärung* H *für Daten* D *ist umso schöner, je kleiner* K(H) *ist.*

Man mag Prinzip 2 aber auch anders verstehen, nämlich dahingehend, dass nicht nur die Hypothese selbst möglichst einfach sein soll, sondern auch die Art und Weise, in der sie die Daten erklärt. Beispielsweise mag der Fall der Rue Morgue durch einen wahnsinnigen Nachbarn der ermordeten Frauen erklärt werden. Das ist eine recht einfache Hypothese, aber die Art und Weise, in der sie den Fall erklären würde, müsste wohl recht kompliziert ausfallen. Beispielsweise scheint, den Rue

8 Eine Turing Maschine $M$ ist präfixfrei, wenn für je zwei verschiedene binäre Strings $x$ und $y$, so dass $M$ sowohl bei Input $x$ als auch bei Input $y$ hält, gilt, dass weder $x$ ein Präfix von $y$ noch $y$ ein Präfix von $x$ ist.

Morgue Daten zufolge, der Mörder über übermenschliche Kräfte zu verfügen und irgendwie muss die Hypothese mit dem wahnsinnigen Nachbarn damit zurechtkommen. Natürlich ist das irgendwie möglich, wenn man unbedingt will, aber vermutlich nur sehr umständlich. Es wäre ein sicherlich langwieriges und wahrscheinlich recht künstliches Unterfangen, die Daten der Rue Morgue mit der Hypothese des wahnsinnigen Nachbarn zu erklären. In diesem Sinne mag man diese Hypothese nicht als eine einfache akzeptieren.

Die Forderung, dass die Art und Weise, in der eine Hypothese H die Daten D erklärt, möglichst einfach sein soll, kann man dadurch formalisieren, dass man verlangt, dass die Beschreibung der Daten D mittels der Hypothese H möglichst einfach ist, also dass K(D|H) möglichst klein ist. Wir erhalten folgende

**Zweite Formalisierung von Prinzip 2**
*Eine Erklärung* H *für Daten* D *ist umso schöner, je kleiner* K(H) + K(D|H) *ist.*

## 4. Schlussfolgerungen.

Was heißt es, dass eine Hypothese bestimmte Daten ‚erklärt'? Wie in der Einleitung gesagt, wollen wir dieser Frage aus dem Weg gehen und nur zwei Fälle unterscheiden, nämlich je nachdem ob man einen *deterministischen Standpunkt* oder einen *probabilistischen Standpunkt* bezieht. Wir werden an Ort und Stelle erläutern, was wir damit meinen. Unsere Ziele sind die folgenden beiden Schlussfolgerungen:

**Erste Schlussfolgerung**
*Vom deterministischen Standpunkt aus gesehen ist Prinzip 1 äquivalent zu Prinzip 2 in dessen erster Formalisierung.*

**Zweite Schlussfolgerung**
*Vom probabilistischen Standpunkt aus gesehen ist Prinzip 1 äquivalent zu Prinzip 2 in dessen zweiter Formalisierung.*

Die erste Schlussfolgerung (Abschnitt 4.2) erhalten wir mittels eines fundamentalen Resultats der algorithmischen Informationstheorie, nämlich dem Kodierungstheorem. Wir stellen dieses Theorem im folgenden Abschnitt vor. Um die zweite Schlussfolgerung (Abschnitt 4.4) zu gewinnen, benötigen wir den Begriff der Martin-Löf Zufälligkeit. Wir stellen diesen Begriff skizzenhaft in Abschnitt 4.3 vor.

### 4.1. Das Kodierungstheorem.

Das Kodierungstheorem besagt, dass ein binärer String umso einfacher ist, je größer seine universale Wahrscheinlichkeit ist (und umgekehrt). Genauer gesagt:

**Theorem 2**
$\mathrm{K}(x) = -\log \mathbf{m}[x] \pm O(1)$.

Einige Erklärungen dazu: die Notation $\pm\, O(1)$ soll bedeuten, dass die Gleichheit nur plus/minus höchstens einer Konstanten gilt. Mit anderen Worten: $-\log \mathbf{m}[x] - \mathit{const} \leq \mathrm{K}(x) \leq -\log \mathbf{m}[x] + \mathit{const}$ für eine geeignete positive Konstante *const*. Mit log bezeichnen wir den Logarithmus zur Basis 2. Wir können Theorem 2 also auch so schreiben: $1/2^{\,\mathrm{K}(x)\, \pm\, O(1)} = \mathbf{m}[x]$. Theorem 2 lässt sich derart verstehen, dass man ein universales (im Sinne von Theorem 1) effektives Wahrscheinlichkeitsmaß bekommt, wenn man jedem binären String *x* die Wahrscheinlichkeit $1/2^{\mathrm{K}(x)}$ zuordnet[9].

---

9 Dass dies ein effektives Wahrscheinlichkeitsmaß definiert, ist zwar nicht offensichtlich, aber wahr (s. Anm. 8).

## 4.2. Die ästhetischen Prinzipien von einem deterministischen Standpunkt aus gesehen.

Was ist eine Erklärung? Der erste Fall, den wir betrachten, besteht darin, einen *deterministischen Standpunkt* zu beziehen. Das soll heissen zu fordern, dass

(2) $\Pr[D|H] = 1$.

Das kann man so verstehen, dass, *wenn* die Hypothese H wahr ist, dann *muss* der Fall D eintreten (mit Wahrscheinlichkeit 1). Der deterministische Standpunkt verlangt etwa eine kausale Ursache oder etwas ähnlich Zwingendes als Erklärung. In der Sprechweise von Abschnitt 2.1 besteht der deterministische Standpunkt darin, maximalen Erklärungswert zu verlangen.

Wir begründen nun die erste Schlussfolgerung: Nach der Formalisierung von Prinzip 1 ist eine Erklärung H für Daten D umso schöner, je größer $\Pr[D|H] \cdot \mathbf{m}[H]$ ist. Gemäß dem deterministischen Standpunkt (2) heißt das: je größer $\mathbf{m}[H]$ ist, das heißt, je kleiner $-\log \mathbf{m}[H]$ ist. Gemäß Theorem 2 heißt das, je kleiner K(H) ist. Und das ist die erste Formalisierung von Prinzip 2.

## 4.3. Martin-Löf Zufälligkeit.

Vielleicht ist der deterministische Standpunkt zu restriktiv und es ist übertrieben, einen maximalen Erklärungswert zu fordern. Eine bestimmte Position, die davon abrückt, wollen wir den probabilistischen Standpunkt nennen und im Folgenden beschreiben.
Man mag für die Daten, dass der Fußboden nass ist, die Erklärung „Ich habe Wasser verschüttet" gelten lassen, obwohl natürlich nicht ausgeschlossen ist, dass der Fußboden (etwa mittels technisch versiertem Eingreifen von Dänikens Außerirdischen) mein Wasserverschütten trocken übersteht. In den wohl meisten Fällen sind derartige Einwände absurd, denn *typischerweise* wird der Fußboden durch das Verschütten von Wasser eben nass und deswegen ist die gegebene Erklärung ausreichend. Aber was heißt es, dass die beobachteten Daten (der nasse Fußboden) sehr ‚typische' Daten sind *gegeben* die Hypothese (Wasserverschütten)?
Die Frage nach der Bedeutung des Begriffs ‚typisch' ist Jahrhunderte alt und wurde in der algorithmischen Informationstheorie von Martin-Löf zufriedenstellend beantwortet. Hier ist nicht der Platz, um diese Antwort darzustellen. Wir geben hier nur eine Definition und einige Kommentare dazu, dass sie tatsächlich ‚funktioniert'.[10]

Für ein gegebenes effektives Wahrscheinlichkeitsmaß Pr wollen wir erklären, was ein ‚für Pr typischer' binärer String ist. Für eine mehr oder weniger beliebig festgelegte positive Konstante *const* definieren wir: Ein binärer String *x* ist *typisch für* Pr (man sagt auch *Martin-Löf zufällig bezüglich* Pr) genau dann, wenn

$$\log (\mathbf{m}[x] / \Pr[x]) \leq const.$$

Das kann man so verstehen, dass die Wahrscheinlichkeit von *x* gemäß Pr nicht zu sehr von seiner universalen Wahrscheinlichkeit abweichen soll. Es lässt sich zeigen, dass die in diesem Sinne für Pr typischen *x* gerade die sind, die sich durch keine Besonderheit auszeichnen in dem Sinne, dass sie zu allen (in geeignetem Sinne effektiven) Mehrheiten (im Sinne von Pr) gehören. Außerdem kann man (formal sauber ausdrücken und dann) zeigen, dass die meisten (im Sinne von Pr) binären Strings typisch für Pr sind.

Mit diesem Begriff können wir den *probabilistischen Standpunkt* so verstehen: Eine Hypothese H ist eine Erklärung von Daten D nur, wenn D typisch ist für $\Pr[\,\cdot\,|H]$ (also für Pr gegeben die Hypothese H), das heißt, nur wenn

10 Der interessierte Leser sei verwiesen auf Calude (Anm. 4) und Vitanyi, Li: Introduction (Anm. 4), einen schnellen Einblick gewährt Vitanyi, Lie: Description (Anm. 4), Appendix C.

(3) $\log (\mathbf{m}[D|H] / Pr[D |H]) \leq \textit{const}$.

Anders ausgedrückt, verlangt der probabilistische Standpunkt von einer Hypothese, die als Erklärung für die Daten D gelten soll, dass der Erklärungswert ungefähr der universalen Wahrscheinlichkeit der Daten gegeben die Hypothese gleicht.

Es ist leicht einzusehen, dass der probabilistische Standpunkt eine Lockerung des deterministischen darstellt: Aus Gleichung (2) folgt die Ungleichung (3) (wie auch immer *const* gewählt ist). Eine Hypothese, die vom deterministischen Standpunkt aus gesehen eine Erklärung ist, ist also auch vom probabilistischen Standpunkt aus gesehen eine Erklärung.

## 4.4. Die ästhetischen Prinzipien von einem probabilistischen Standpunkt aus gesehen.

Wir beurteilen die beiden Prinzipien jetzt von einem probabilistischen Standpunkt aus, das heißt, wir nehmen an, dass für eine Erklärung H von D die Ungleichung (3) gilt.
Zuerst eine Beobachtung: Aus (3) und der Universalität von **m** (siehe Theorem 1) folgt leicht, dass $\log (\mathbf{m}[D|H] / Pr[D|H]) = 0 \pm O(1)$. Gemäß den Rechenregeln für den Logarithmus gilt somit

(4) $\log \mathbf{m}[D|H] = \log Pr[D|H] \pm O(1)$.

Wir begründen jetzt die zweite Schlussfolgerung: Gemäß der Formalisierung von Prinzip 1 ist eine Erklärung H für D umso schöner, je größer $Pr[D|H] \cdot \mathbf{m}[H]$ ist, das heißt, je kleiner $(-\log Pr[D|H] - \log \mathbf{m}[H])$ ist. Nach (4) heißt das, je kleiner $(-\log \mathbf{m}[D|H] - \log \mathbf{m}[H])$ ist. Nach Theorem 2 heißt das, je kleiner $K(D|H) + K(H)$ ist. Und das ist die zweite Formalisierung von Prinzip 2.

## *5. Nachbemerkung.*

In der Philosophie angelsächsischer Tradition haben mit dem Aufkommen der modernen Logik nach Frege Formalisierungen einen hohen methodischen Stellenwert bekommen und die analytische Philosophie ist enstanden. Ähnlich wie eine analytische Philosophie ließe sich auch eine analytische Literaturwissenschaft vorstellen.
Die Methodik der analytischen Philosophie begann mit Bertrand Russell und findet sich vor allem im Wiener Kreis, dort vor allem bei Rudolf Carnap. Ich denke, dass so die Philosophie viel an Klarheit gewonnen hat. Wittgensteins Diktum, dass sich alles, was sich sagen lasse, auch klar sagen lasse, mag als Sinnkriterium fraglich sein, mag auch selbst unklar sein, als methodischer Imperativ (gegen den Strich) gelesen ist es aber wohl so etwas wie eine Minimalanforderung an eine Konversation und ausserdem gut gegen Opium. Erinnern wir uns an Carnaps Toleranzprinzip:

> In der Logik gibt es keine Moral. Jeder mag seine Logik, d.h. seine Sprachform, aufbauen wie er will. Nur muss er, wenn er mit uns diskutieren will, deutlich angeben, wie er es machen will, syntaktische Bestimmungen geben anstatt philosophischer Erörterungen.[11]

Wir haben zwei ästhetische Prinzipien formalisiert und so gewisse Einsichten über deren Verhältnis hergeleitet. Ich weiß nicht, ob diese Einsichten aus literaturwissenschaftlicher Sicht langweilig oder gar falsch sind. Ich glaube aber, daß die Methode grundsätzlich eine gute ist.

---

11 Rudolf Carnap: Logische Syntax der Sprache, in: Ders.: Schriften zur wissenschaftlichen Weltauffassung, hrsg. von Philipp Frank und Moritz Schlick, Bd. 8, Wien 1934, S. 45.