Novelty, Information and Surprise
Novelty, Information and Surprise
Contents

Introduction ................................................................. xiii
References ........................................................................... xxii

Part I Surprise and Information of Descriptions

1 Prerequisites from Logic and Probability Theory ................. 3
  1.1 Logic and Probability of Propositions ......................... 3
  1.2 Mappings, Functions and Random Variables ............... 5
  1.3 Measurability, Random Variables, and Expectation Value 7
  1.4 Technical Comments ........................................... 10
  References .................................................................. 10

2 Improbability and Novelty of Descriptions ......................... 11
  2.1 Introductory Examples ......................................... 11
  2.2 Definition and Properties ..................................... 14
  2.3 Descriptions .................................................... 15
  2.4 Properties of Descriptions .................................... 18
  2.5 Information and Surprise of Descriptions ................. 24
  2.6 Information and Surprise of a Random Variable .......... 30
  2.7 Technical Comments ......................................... 31
  2.8 Exercises ....................................................... 32
  References .................................................................. 34

3 Conditional and Subjective Novelty and Information .......... 35
  3.1 Introductory Examples ......................................... 35
  3.2 Subjective Novelty ............................................. 36
  3.3 Conditional Novelty ........................................... 38
  3.4 Information Theory for Random Variables ................. 42
  3.5 Technical Comments ......................................... 44
  3.6 Exercises ........................................................ 45
  References .................................................................. 46
# Part II  Coding and Information Transmission

4  **On Guessing and Coding** ................................................................. 51  
4.1 Introductory Examples ................................................................. 51  
4.2 Guessing Strategies ....................................................................... 53  
4.3 Codes and Their Relation to Guessing Strategies ......................... 54  
4.4 Kraft’s Theorem ............................................................................. 56  
4.5 Huffman Codes ............................................................................... 57  
4.6 Relation Between Codewordlength and Information ...................... 58  
4.7 Technical Comments ..................................................................... 60  
4.8 Exercises ......................................................................................... 60  
References ............................................................................................. 62

5  **Information Transmission** .......................................................... 63  
5.1 Introductory Examples ................................................................. 63  
5.2 Transition Probability ...................................................................... 65  
5.3 Transmission of Information Across Simple Channels .................. 67  
5.4 Technical Comments ..................................................................... 71  
5.5 Exercises ......................................................................................... 72  
Reference .............................................................................................. 74

# Part III  Information Rate and Channel Capacity

6  **Stationary Processes and Their Information Rate** ....................... 77  
6.1 Introductory Examples ................................................................. 77  
6.2 Definition and Properties of Stochastic Processes ......................... 78  
6.3 The Weak Law of Large Numbers ............................................... 80  
6.4 Information Rate of Stationary Processes ..................................... 81  
6.5 Transinformation Rate ..................................................................... 84  
6.6 Asymptotic Equipartition Property .............................................. 85  
6.7 Technical Comments ..................................................................... 87  
6.8 Exercises ......................................................................................... 87  
References ............................................................................................. 88

7  **Channel Capacity** ........................................................................... 89  
7.1 Information Channels ..................................................................... 89  
7.2 Memory and Anticipation ............................................................ 90  
7.3 Channel Capacity ........................................................................... 91  
7.4 Technical Comments ..................................................................... 94  
7.5 Exercises ......................................................................................... 94  
References ............................................................................................. 95

8  **How to Transmit Information Reliably with Unreliable Elements (Shannon’s Theorem)** ......................................................... 97  
8.1 The Problem of Adapting a Source to a Channel ........................... 97  
8.2 Shannon’s Theorem ....................................................................... 98
8.3 Technical Comments ................................................. 101
8.4 Exercises .............................................................. 101
References..................................................................... 101

**Part IV  Repertoires and Covers**

9  Repertoires and Descriptions .............................................. 105
9.1 Introductory Examples ............................................... 106
9.2 Repertoires and Their Relation to Descriptions ................. 109
9.3 Tight Repertoires ..................................................... 115
9.4 Narrow and Shallow Covers ......................................... 117
9.5 Technical Comments .................................................. 119
9.6 Exercises .............................................................. 120
References..................................................................... 120

10 Novelty, Information and Surprise of Repertoires .............. 123
10.1 Introductory Examples ............................................... 123
10.2 Definitions and Properties ......................................... 125
10.3 Finding Descriptions with Minimal Information ................ 133
10.4 Technical Comments ................................................ 138
10.5 Exercises .............................................................. 138
References..................................................................... 139

11 Conditioning, Mutual Information, and Information Gain ..... 141
11.1 Introductory Examples ............................................... 141
11.2 Conditional Information and Mutual Information .............. 142
11.3 Information Gain, Novelty Gain, and Surprise Loss .......... 146
11.4 Conditional Information of Continuous Random Variables ... 152
11.5 Technical Comments ................................................ 154
11.6 Applications in Pattern Recognition, Machine Learning, and Life-Science ........................................ 155
11.7 Exercises .............................................................. 156
References..................................................................... 157

**Part V  Information, Novelty and Surprise in Science**

12 Information, Novelty, and Surprise in Brain Theory ........... 161
12.1 Understanding Brains in Terms of Processing and Transmission of Information ........................................ 161
12.2 Neural Repertoires ................................................... 166
12.3 Experimental Repertoires in Neuroscience ...................... 167
12.3.1 The Burst Repertoire ........................................... 168
12.3.2 The Pause Repertoire .......................................... 170
12.3.3 The Coincidence Repertoire .................................. 170
12.3.4 The Depolarization Repertoire ............................... 173
12.4 Neural Population Repertoires: Semantics and Syntax ....... 173
12.5 Conclusion ............................................................ 175
12.6 Technical Comments ................................................. 175
12.6.1 Coincidence ................................................. 179
12.6.2 Coincidental Patterns ....................................... 179
12.6.3 Spatio-Temporal Patterns ................................... 179
References .................................................................... 181

13 Surprise from Repetitions and Combination of Surprises ......... 189
13.1 Combination of Surprises ............................................ 189
13.2 Surprise of Repetitions ............................................... 191
13.3 Technical Comments ................................................. 194
References .................................................................... 194

14 Entropy in Physics .......................................................... 195
14.1 Classical Entropy ..................................................... 195
14.2 Modern Entropies and the Second Law ............................. 198
14.3 The Second Law in Terms of Information Gain .................... 201
14.4 Technical Comments ................................................. 204
References .................................................................... 204

Part VI Generalized Information Theory

15 Order- and Lattice-Structures ............................................ 207
15.1 Definitions and Properties ........................................... 207
15.2 The Lattice \( \mathcal{D} \) of Descriptions ................................ 213
15.3 Technical Comments ................................................. 214
Reference ..................................................................... 215

16 Three Orderings on Repertoires .......................................... 217
16.1 Definition and Basic Properties ................................. 217
16.2 Equivalence Relations Defined by the Orderings .............. 220
16.3 The Joins and Meets for the Orderings ......................... 222
16.4 The Orderings on Templates and Flat Covers ................... 226
16.5 Technical Comments ................................................. 227
16.6 Exercises ................................................................... 228
References .................................................................... 228

17 Information Theory on Lattices of Covers ......................... 229
17.1 The Lattice \( \mathcal{C} \) of Covers ...................................... 229
17.2 The Lattice \( \mathfrak{C} \) of Finite Flat Covers ......................... 231
17.3 The Lattice \( \mathfrak{R} \) of (Clean) Repertoires ....................... 232
17.4 The Lattice \( \mathfrak{E} \) of Templates ................................. 233
17.5 The Lattice \( \mathfrak{P} \) of Partitions .................................. 234
17.6 Technical Comments ................................................. 235
17.7 Exercises ................................................................... 235
References .................................................................... 235
List of Figures

Fig. 1  Examples of covers (top) and the induced hierarchical structures (bottom) ............................................................... xv
Fig. 2.1 Model for descriptions of events .................................................. 16
Fig. 2.2 Example of propositions about positions on a table ......................... 17
Fig. 2.3 Plot of function $h(p)$ ................................................................. 25
Fig. 2.4 Plot of function $I(p)$ ................................................................. 28
Fig. 4.1 Example of a question strategy .................................................. 52
Fig. 4.2 Example of a question strategy .................................................. 52
Fig. 4.3 Tree picturing a question strategy .............................................. 54
Fig. 5.1 An information channel ............................................................... 66
Fig. 5.2 Channel example ................................................................. 67
Fig. 5.3 Three different channels ............................................................. 73
Fig. 6.1 Channel example ................................................................. 78
Fig. 7.1 Data transmission on a channel .................................................. 90
Fig. 7.2 Two channels ................................................................. 94
Fig. 7.3 Two simple channels ................................................................. 95
Fig. 8.1 Channel and stationary process .................................................. 98
Fig. 9.1 Illustration of tightness .............................................................. 112
Fig. 9.2 Illustration of cleanness. The vertically hatched region on the left can be removed, because it is the union of the two diagonally hatched regions ............................................ 113
Fig. 9.3 Illustration of the product of two repertoires .................................. 116
Fig. 9.4 Examples of narrow covers and chains ........................................ 118
Fig. 9.5 Example for a shallow repertoire (a) and for a chain (b) .................... 118
Fig. 9.6 Illustration of classes of repertoires ............................................ 119
Fig. 12.1 A single neuron .............................................................. 167
Fig. 12.2  Burst novelty as a function of time for 3 individual spike-trains and a simulated Poisson-spike-train .................. 169
Fig. 12.3  Burst surprise as a function of burst novelty ...................... 170
Fig. 12.4  Histogram of novelty values of spike-bursts. Novelty based on Poisson distribution. Spikes from visual cortex neurons in awake behaving cats. For more details see Legéndy and Salcman (1985) from where Fig. 12.4 and 12.5 are adapted. ...................................... 177
Fig. 12.5  Spike-burst statistics. a) Histogram of the spike rate during high surprise bursts (thick bars: \( N > 10 \), thin bars: \( N > 20 \)), b) Histogram of the number of spikes in high surprise bursts. Preparation as in Fig. 12.4 from Legéndy and Salcman (1985). ....................................... 178
Fig. 16.1  Possible relationships between the orderings \( \leq_1 \), \( \leq_2 \), and \( \leq_3 \) ................................................................. 218
Fig. 16.2  Repertoires illustrating that \( \leq_2 \) is not a lattice ................. 223
Introduction

Nowadays there are many practical applications of information theory in fields like pattern recognition, machine learning, and data mining (e.g., Deco and Obradovic 1996; MacKay 2005), used in particular in the life sciences [e.g., Herzel et al. (1994); Schmitt and Herzel (1997); Bialek et al. (2007); Taylor et al. (2007); Tkačík and Bialek (2007); Koepsell et al. (2009)], i.e., far beyond the classical applications in communication technology. But Claude Shannons groundbreaking original concept of information remained essentially unchanged.

The main purpose of this book is to extend classical information theory to incorporate the subjective element of interestingness, novelty, or surprise. These concepts can only be defined relative to a person’s interests, intentions, or purposes, and indeed classical information theory was often criticized for not being able to incorporate these ideas. Actually, classical information theory comes quite close to this, when introducing the information contained in a proposition or statement $A$ (as $- \log_2 p(A)$). But in everyday life most commonly this information is not really transferred if a person $X$ tells the statement $A$ to another person $Y$, for example because $Y$ may not be interested in $A$ or because $Y$ may rather be interested in the fact that it is $X$ who tells $A$. An interesting extension of information theory could consider the following question: If $Y$ is interested in $B$ instead of $A$ and perhaps $B$ largely overlaps with $A$, then how much information does $Y$ obtain from being told $A$? This question and other similar ones will be answered in this book; they are not totally alien to classical information theory. This means that our new theory does not have to go far from it. In a technical sense it can be regarded as only a slight (but perhaps important) extension of Shannon’s definition of information from partition to covers.

This needs some explanation: Shannon tried to define his information as an objective almost physical quantity (measured in bit). This led him to define information for random variables $X$. For a discrete random variable $X$ with finitely many possible values $x_1, x_2, \ldots, x_n$ he defined $I(X) = - \sum_i p[X = x_i] \log_2 p[X = x_i]$, i.e., as the average of $- \log_2 p[X = x_i]$ where the possible outcomes $[X = x_i]$ of $X$ are now the propositions $A$ of interest. Again, this definition presupposes that we are equally interested in all outcomes $x_i$ of the random variable. This may not
always be the case. For example, if we consider a roulette game represented by a random variable $X$ with possible values in $\{0, 1, 2, \ldots, 36\}$, then a person who has put money on the number 13, the row 31, 32, 33, and the odd numbers, certainly wants to know the outcome number, but he will be more interested in the numbers mentioned above. The new concept of novelty defined here takes care of this aspect of information which goes slightly beyond classical information theory.

Shannon’s definition uses the statements $[X = x_i]$. In classical probability theory these are called “events” and modeled as subsets of a universal set $\Omega$ of all possible events. These sets actually form a partition of $\Omega$, meaning that they are mutually exclusive and cover all of $\Omega$. A cover is a more general set of “events” than a partition, where mutual exclusiveness is not required. My idea, the foundation of this theory, is to use a cover (instead of a partition) to model the set of all propositions or statements a person is interested in, and to define information for such covers.

Starting from a definition of information on partitions the step to covers appears as a rather straightforward generalization, namely omitting the requirement of mutual exclusiveness or disjointness of the propositions. Technically, partitions have a number of very useful properties which seem to be necessary to prove even the most elementary theorems of information theory (for example, the monotonicity and subadditivity of information). So the main body of this book is devoted to the development of a workable extension of classical information theory to (possibly) overlapping sets of propositions called repertoires or covers, which is the theory of novelty and which coincides with classical information theory when the covers are partitions.

This move from partitions to covers allows to turn our attention to the rich logical structures that are possible between the propositions in arbitrary covers. In general, these structures can be described most adequately as hierarchies (see Fig. 1). This turn to the possible logical structures that may underlie definitions of novelty, information, or even physical entropy provides a new perspective for the interpretation of these concepts, for example in thermodynamics and in neuroscience. This may be more a philosophical point, but it can have practical implications (and it was one of my strongest motivations to write this book). More about it can be found in Part V.

When we consider arbitrary covers $\alpha = \{A_1, \ldots, A_n\}$ of a probability space $\Omega$ instead of partitions, it actually makes sense to distinguish different information-like concepts, which we call information, novelty, and surprise. In a nutshell, the information of $\alpha$ is the minimum average number of yes–no questions that we need to determine for every $\omega \in \Omega$ one element $A \in \alpha$ that describes it (i.e., $\omega \in A$)—the information needed for $\alpha$.

The novelty of $\alpha$ is the average maximum information $I(A) = -\log_2 p(A)$ that we can obtain for $\omega \in \Omega$ from an $A \in \alpha$ that describes it—the novelty obtained from $\alpha$. This means the novelty for $\omega$ is $N(\omega) = \max\{-\log_2 p(A); \omega \in A \in \alpha\}$.

The surprise of $\alpha$ has a more statistical flavor: if we have a very rich cover $\alpha$, in the extreme case $\alpha$ may contain all singletons $\{\omega\}$ ($\omega \in \Omega$), then we always obtain a lot of novelty, but this is not really surprising. So we say that we get really surprised when observing $\omega \in \Omega$ through $\alpha$, when $N(\omega)$ is large compared to other values
This leads to the definition of surprise as

\[ S(\omega) = - \log_2 p[N \geq N(\omega)] = - \log_2 p(\{\omega' \in \Omega: N(\omega') \geq N(\omega)\}). \]

Besides introducing the new concepts of novelty and surprise, this book also extends classical Shannon information. Therefore one may hope that this new more general treatment of information can be used to solve some problems that could not be solved with classical information theory. This is indeed the case when we consider problems which need overlapping propositions for their formulation. Let me give one example for this here, which is treated more extensively in Chap. 10 and which appears rather innocent on the surface. You are observing a shell game where there are eight shells on the table and there is a coin (say 1 €) under two of them. How much information do you get, when somebody points to one of the shells and tells you that it contains a coin? How would you devise an optimal guessing strategy that determines the position of one of the two coins, and how many questions do you need on average?

You cannot answer these questions properly only with classical information theory and the correct answers are given in Chap. 10.

**Organization of the Book**

The book is organized into 6 parts.

*Part I:* Introduces the basic concepts on an elementary level. It contains a brief introduction to probability theory and the new concept of a
description, which maps $\Omega$ into $\Sigma$ for a probability space $(\Omega, \Sigma, p)$, i.e., it associates with every elementary event $\omega \in \Omega$ a proposition $A \in \Sigma$ describing it. Already on the level of descriptions we can distinguish information, novelty, and surprise.

**Part II:** Recapitulates classical coding theory. It is not essential for the new concepts developed in this book, but for students of information theory it is necessary to understand the practical meaning of information and the related notions defined here.

**Part III:** Introduces some more background on stochastic processes which is necessary to prove Shannon’s classical theorem, the backbone of information theory. The material in this part is not new, but in my opinion some proofs become a bit easier due to the new concept of description.

**Part IV:** Contains the core ideas of this book. It defines various structures on the set of all covers and motivates their introduction by various practical examples. It also contains the definitions of information, novelty, and surprise for covers or repertoires and shows how these numbers can be calculated in practice.

**Part V:** Shows some applications of our new and more general view of information theory in neuroscience, brain theory, and the physics of entropy.

**Part VI:** Concentrates on the mathematical structures on which the new generalized information theory is built. It harvests and combines the mathematical results obtained in previous parts (mainly in Part IV). It defines six new and interesting lattices of covers which could become the subject of further mathematical investigations.

This book contains a mathematical theory of information, novelty, and surprise. It should, however, be readable for everyone with a basic mathematical background (as given in the first year of most scientific curricula) and the patience and stamina necessary to follow through a mathematical kind of exposition and to do at least some of the exercises.

**Philosophy of the Book**

The book provides a brief and comprehensive exposition of classical information theory, from a slightly unusual point of view, together with the development of a new theory of novelty and surprise. These new concepts are defined together with the concept of information as a complementary companion. The word surprise invokes many mostly subjective notions and this subjectivity is brought in on purpose to complement the seemingly more “objective” notion of information.

The “subjective vs. objective” debate has a long history in the field of probability theory and statistics. There it focuses on the question of the nature or the origin of probability, e.g., Keynes 1921; Jeffreys 1939; Kerridge 1961; de Finetti 1974;
Jeffrey 1992, 2004. This discussion can be and has been carried over to information theory, but this is not the purpose of this book. Instead I have discovered an additional source of subjectivity in probability and information theory that is perhaps less relevant in probability but has a definite impact on information theory once it is brought into the focus of attention. Classical information theory tried to determine the amount of information contained in a message (as a real number measured in bits). To do this it has to presuppose that the message is exchanged between two agents (that are normally assumed to be people not animals, computers, or brain regions), who have already agreed on a common language for the expression of the message. This approach makes it possible to develop information theory in a discrete framework, dealing mainly with a finite alphabet from which the messages are composed. Thus classical information theory does not consider the process of perception or of formation of the messages. It starts where this process has ended.

I believe and I will actually show in this book that information theory needs only a slight modification to include this process of perception and formation of a message. This process is captured in the definition of a description which connects every event \( \omega \) that can happen, with a proposition \( d(\omega) \) about it. When we describe actual events that have happened, we will normally be unable to give an exact account of them, let alone of the total state of the world around us. We are restricted both by our senses and by our language: we cannot sense everything that there is and we cannot express everything that we sense (at least not exactly). So the description \( d(x) \) that we give about \( x \) will usually be true not only for \( x \) but also for other events \( y \) which are somewhat similar to \( x \).

Now it is quite clear that classical information theory does not deal with events \( x \) that happen, but rather with descriptions of these events or propositions about those events. Unfortunately, this simple fact is obscured by the usual parlance in probability theory and statistics where a proposition is called an “event” and an event is called an “elementary event.” This leads to such strange phrases as “\( \Omega \) is the certain event” and “\( \emptyset \) is the impossible event.” What is meant is that \( \Omega \) is the trivial proposition, which is true because it says nothing about the event \( x \), and that \( \emptyset \) is never true because it is a self-contradictory statement (like “\( A \) and not \( A \)”) about the event \( x \). Due to this strange use of the language (which is usually carried over from probability theory to information theory) it may easily happen that this humanoid or language-related aspect of the concept of information is forgotten in practical applications of information theory. This is harmless when information theory is used to optimize telephone lines, but it may become problematic when information theory is applied in the natural sciences. This had already happened in the nineteenth century when the term entropy was introduced in statistical mechanics to explain the second law of thermodynamics and it is happening again today when information theory is used in theoretical neuroscience or in genetics, e.g., Küppers 1986. One problem is that in such applications the concept of information may be taken to be more “objective” than it actually is. This argument eventually has led me and also others to some rather subtle and probably controversial criticisms of these applications of information theory Knoblauch and Palm 2004; Palm 1985, 1996; Bar-Hillel and Carnap 1953 which really were not the main motive for writing this
book and which are neither part of the theory of novelty and surprise nor its most interesting applications. In a nutshell, the situation can be described as follows. Classical Shannon information theory requires an agreement between the sender and the receiver of the messages whose information content has to be determined. So information travels from a human sender to a human receiver, and the agreement concerns the code, i.e., the propositional meaning of the symbols that constitute the message; this implies that both sender and receiver use the same description of events. The slightly broader information theory developed here includes the situation where information about a purely physical observation is extracted by human observation, so only the receiver needs to be a human. In many scientific and everyday uses of information terminology, however, neither the sender nor the receiver is human. And the problem is not merely that instead of the human we have some other intelligent or intentional being, for example an alien, a monkey, a chess-computer, or a frog. Information terminology is also used for completely “mechanical” situations. For example, a cable in my car carries the information that I have set the indicator to turn left to the corresponding lights. Or a cable transmits visual information from a camera to the computer of my home intrusion-warning system.

In biology and in particular in neuroscience this common use of information terminology may interfere in strange ways with our ontological prejudices, for example concerning the consciousness of animals, because on the one hand information terminology is handy, but on the other hand we don’t want to imply that the receiver of the information has the corresponding properties comparable to a human. For example, we may want to quantify the amount of visual information the optic nerve sends to the brain of a frog (Letvin et al. 1959; Atick 1992), without assuming that the frog (let alone its brain) has made some agreement with its eyes about the meaning of the signals that are sent through the nerve. Similarly, we can easily classify the maximal amount of information that can be expressed by our genes, but we get into much deeper waters when we try to estimate how much information is actually transmitted, and who is the sender and who is the receiver. Does father Drosophila transmit some information (for example, about how to behave) to his son by his genes? Or is it somehow the whole process of evolution that produces information (Küppers 1986, see also Taylor et al. 2007) and who is the receiver of this information? The usual way out of this dilemma is to avoid the question who might be the sender and who might be the receiver altogether. In the technical examples eventually both are always humans, anyway. In the case of information transmission by genes, neurons, or the optic nerve one can argue that we are just interested in the physical properties of the “device” that limit the amount of transmittable information, i.e., the channel capacity. In all these three cases actually, the channel capacity has been estimated shortly after its definition by Shannon 1948. In the case of the genome, the capacity is simply twice the number of base pairs of the DNA-molecule. But this leads to the question how much of this capacity is actually used and, invariably in biology, to the suspicion that it is much less than the capacity. So it is found out that most of the genetic information is “not used” or “not coding”, or that most of the visual information available from our
optic nerve is not “consciously perceived,” neither by us nor by our experimental animals. Somehow people don’t easily accept such statements perhaps because they seem to conflict with the idea that animals are well designed. Perhaps these observations again reflect the original “dilemma of the receiver” which we tried to circumvent with the rather safe capacity argument. My position is that, at least in neuroscience, the more subjective turn to information theory presented here may help to alleviate this problem and to find better estimates of transinformation (or rather mutual novelty) flows in neural systems that are in between the two extremes of neurophysiological channel capacity (e.g., MacKay and McCulloch 1952) and transinformation from experimental stimuli to behavioral responses (e.g., Eckhorn et al. 1976; Borst and Theunissen 1999) because received novelty < transmitted information < information capacity. Chapter 12 contains some ideas and first results in this direction.

The new concepts of novelty and surprise may also help to capture some puzzling aspects of everyday use of information that are hard to understand in terms of classical information theory. In many expositions of mathematical information theory it is obvious from the beginning that information in fact deals with propositions. This is the case when information is primarily defined for partitions, i.e., for complete sets of mutually exclusive propositions (about the event \( \omega \)). Although I sympathize with this approach, I believe that this way of defining information is still too narrow, because it does not allow for mutually overlapping propositions. Indeed, I think that in everyday life we usually work with partially overlapping concepts which cannot be sharply separated from each other, or even with hierarchies of concepts which contain each other. Why should it not also be possible to allow such propositions in information theory?

Such a more general approach makes it possible to understand the use of the term novelty or surprise in everyday language where it seems to be unjustified from the point of view of statistics or classical information theory. For example, if in the state lottery the numbers (1, 2, 3, 4, 5, 6) were drawn we would be much more surprised than by the numbers (5, 11, 19, 26, 34, 41), although, of course, both sequences should have exactly the same small chance of being drawn. The reason for our surprise in the first case seems to be that this sequence can be exactly described in a very simple way: it consists of the first six numbers. On the other hand there seems to be no simple description for the second sequence. To remember it one really would have to memorize all the six numbers. Now it is much more probable to obtain a sequence of numbers in the lottery that does not admit a simple exact description than to obtain a sequence like (1, 2, 3, 4, 5, 6) that does. In the special case of (1, 2, 3, 4, 5, 6) we could argue for example that there are only two such extremely simple sequences, namely the last 6 and the first 6 numbers. Of course, there are various degrees of simplicity and our surprise will vary accordingly. For example, the sequence (22, 23, 24, 25, 26, 27) is simple because it is a sequence of consecutive numbers. Therefore it is also very surprising, but less surprising than (1, 2, 3, 4, 5, 6) because there are 44 such sequences including the two very simple ones. A mathematician may even find the sequence (5, 11, 19, 26, 34, 41) a little surprising, because it contains 4 prime numbers. But since there are many possible
sequences containing 4 prime numbers (many more than 44, but much less than all possible sequences), his surprise will certainly be not as large as for the sequence \((22, 23, 24, 25, 26, 27)\).

But if we combine all these different reasons for finding surprise, we may eventually find something surprising in almost every sequence. In this case, it would seem naive to add all surprises we can get for a given sequence from various considerations, rather one would believe that the “real” surprise provided by any concrete sequence becomes less when everything is considered as surprising. This obviously confused wording leads to the distinction between *novelty* and *surprise* that is also made in this book.

**Personal History of the Book**

The idea of writing this book first occurred to me in 1984, when I received an invitation to stay for a year in Berlin as a fellow of the “Wissenschaftskolleg.” Then I thought I could use the freedom gained by the detachment from my daily duties and the secretarial help offered by the Wissenschaftskolleg to produce a manuscript. A few years before that, prompted by my attempts to understand an earlier article by my friend and colleague, Charles Legendy (Legéndy 1975, see also Legéndy and Salcman 1985), I had started to investigate a broader definition of information (Palm 1981). Charles had tried to cast his ideas in brain theory in a framework of information theory; the essential idea was that each neuron in the brain tries to get as much information as possible from the activity of the other neurons and tries to provide as much information as possible for them by its own activity. At that time informational ideas for neuroscience were quite popular in Tübingen, possibly due to the influence of Ernst Pfaffelhuber (Pfaffelhuber 1972). Certainly also my mentor at the MPI in Tübingen, Valentino Braitenberg, often pointed out the importance of information theory (e.g., Braitenberg 1977, 2011), and I needed the essentials for my calculations of associative memory capacity (Palm 1980). It turned out that Charles did not use exactly the usual concept of information as defined in Shannon’s information theory, but rather something very similar, that I had called “surprise” in my article. In this article I already had defined most of the essential concepts of the new theory. So when I started to work on the manuscript in Berlin, my problem seemed to be not whether I would manage to finish it, but whether there was enough material for a whole book.

During the time in Berlin I realized that the ideas that had to be developed actually had an interesting but complicated kinship to the ideas that I had developed in my thesis on topological and measure-theoretical entropy in ergodic theory (see Palm 1975, 1976b). I also realized that they had a bearing on the classical discussions related to physical entropy in thermodynamics and the direction of time which are also related to optimizational ideas based on entropy (Jaynes 1957, 1982). At that time I started to read about Helmholtz and to look into some historical papers by Carnap, Reichenbach, Brillouin, and others. Fortunately, Carl Hempel,
a well-known philosopher of science, was visiting at the same time as a fellow of the “Wissenschaftskolleg” and could help guide my studies in this area.

The year went by very quickly and the manuscript had grown considerably, but there were now more loose ends than before. Back at the Max Planck Institute in Tübingen I managed to use part of my time to complete the first version of the manuscript. I sent it to MIT Press and I was again lucky that the manuscript was seen as potentially interesting although not yet publishable by the reviewers and in particular by Harry Stanton, who encouraged me to keep on working on the manuscript.

In spite of this encouragement, I did not find the time to work on the book for a number of years. Instead I became more deeply involved in brain research and work on methods in neuroscience due to a number of new personal contacts, in particular to Ad Aertsen, Peter Johannesma, and George Gerstein, who came to Tübingen to work in the creative atmosphere in Valentin Braitenberg’s group. I had the possibility of discussing with them (among many other things) my statistical ideas related to the concept of surprise and its possible use in neuroscience. This led to the first appearance of the term surprise in some methodological papers on spike train analysis (Palm et al. 1988; Aertsen et al. 1989) and in the widely used multiunit analysis program by Ad Aertsen. Since that time some of my colleagues (in particular Ad Aertsen and Moshe Abeles) have been pushing me to write these ideas up properly.

After I had left Tübingen in 1988 to become a professor for theoretical brain research at the university of Düsseldorf, I started to use the concepts of description and novelty regularly in the teaching of courses on information theory, first in Düsseldorf and later (after 1991) in Ulm where I became the director of an institute for neural information processing in the computer science department. During the brief period in Düsseldorf, one of my friends from student days in Tübingen, Laura Martignon, joined our group and started to take up work on the book again. She put some of her experience in teaching into the manuscript and helped to make some parts more readable. Later, in Ulm she also taught a course on the subject. Together we submitted the book again to MIT Press and were again encouraged to complete the manuscript. The book seemed to be almost ready for the second time. However, in the next years we both found no time to continue working on it, although since that time I am using it regularly in the teaching of courses on information theory.

Only in the summer of 1997 I did find some time again to work on the book. I partially reorganized it again and isolated the few missing pieces, many of them in the exercises. Fortunately, I found a very talented student, Andreas Knoblauch, who had taken part in one of the courses on information theory and was willing to work on the solutions of many of the exercises. The peace to put it all together and finish most of the remaining work on the manuscript was provided during my sabbatical semester towards the end of 1998 by Robert Miller at the University of Dunedin in New Zealand.

Unfortunately the book was still not completely finished by the end of the millennium. Meanwhile I had more academic and administrative duties as chairman of a collaborative research center and as dean of the computer science department
in Ulm. So it was only towards the end of 2006 when I could take up the book project again. This time I added more motivating examples to the text. And again I found a talented student, Stefan Menz, who integrated everything into a neat LATEX-version of the text. During the last years we also had the opportunity to pursue some of the new ideas and applications of information theory with excellent PhD students in an interdisciplinary school on “Evolution, Information and Complexity,” see Arendt and Schleich (2009).

This book would never have been finished without the help, encouragement, and inspiration from all the people I mentioned and also from many others whom I did not mention. I would like to thank them all!

References


