

Emotion Explained

Edmund T. Rolls

University of Oxford
Department of Experimental Psychology
Oxford
England

OXFORD UNIVERSITY PRESS • OXFORD 2005

Preface

What produces emotions? Why do we have emotions? How do we have emotions? Why do emotional states feel like something? This book seeks explanations of emotion by considering these questions.

One of the distinctive properties of this book is that it develops a conceptual and evolutionary approach (see for example Chapters 2 and 3) to emotion. This approach shows how cognitive states can produce and modulate emotion, and in turn how emotional states can influence cognition. Another distinctive property is that this book links these approaches to studies on the brain, at the level of neuronal neurophysiology, which provides much of the primary data about how the brain operates; but also to neuropsychological studies of patients with brain damage; to functional magnetic resonance imaging (fMRI) (and other neuroimaging) approaches; and to computational neuroscience approaches. The author performs research in all these areas, and this may help the approach to emotion described here to span many levels of investigation. The empirical evidence that is brought to bear is largely from non-human primates and from humans, because of the considerable similarity of their visual and emotional systems associated with the great development of the prefrontal cortex and temporal lobes in primates, and because the overall aim is to understand how emotion is implemented in the human brain, and the disorders that arise after brain damage.

To understand how the brain works, including how it functions in emotion, it is necessary to combine different approaches, including neural computation. Neurophysiology at the single neuron level is needed because this is the level at which information is exchanged between the computing elements of the brain. Evidence from the effects of brain damage, including that available from neuropsychology, is needed to help understand what different parts of the system do, and indeed what each part is necessary for. Neuroimaging is useful to indicate where in the human brain different processes take place, and to show which functions can be dissociated from each other. Knowledge of the biophysical and synaptic properties of neurons is essential to understand how the computing elements of the brain work, and therefore what the building blocks of biologically realistic computational models should be. Knowledge of the anatomical and functional architecture of the cortex is needed to show what types of neuronal network actually perform the computation. And finally the approach of neural computation is needed, as this is required to link together all the empirical evidence to produce an understanding of how the system actually works. This book utilizes evidence from all these disciplines to develop an understanding of how emotion is implemented by processing in the brain.

The overall plan of the book is as follows. Chapter 1 outlines the ways in which this book approaches different types of explanation of emotion, and introduces some of the concepts. Chapter 2 then considers the nature of emotion, producing a theory of emotion, and comparing it to some other theories. Chapter 3 considers the functions of emotion, and leads to a Darwinian theory of the adaptive value of emotion, which helps to illuminate many aspects of brain design and behaviour. Chapter 4 takes the explanation of emotion to the level of how emotion is implemented in the brain. Chapters 5 and 6 extend and complement this

by extending the approach to motivated behaviour in which affect is an important component. In Chapter 5 the motivated behaviour considered is hunger, and in Chapter 6 thirst. Chapter 7 extends the approach to reward and affect produced by brain stimulation, and Chapter 8 to the pharmacology of emotion and addiction. Chapter 9 extends the approach further, to sexual behaviour. Chapter 10 then considers the issue of emotional feelings, which is part of the much larger issue of consciousness. Chapter 11 then synthesizes some of the points made, including how decisions are made and are influenced by emotions. Appendix 1 describes some of the computational framework for understanding how systems in the brain in the form of neural networks perform emotion-related learning. Appendix 2 describes an example of a more detailed neural network approach to emotion-related learning in which the analysis extends from the level of the spiking activity of single neurons up through many levels of investigation to global properties of the system such as the signals measured in functional neuroimaging investigations, and the resulting behaviour. Appendix 3 provides a Glossary of some of the terms. The book thus seeks to explain emotions in terms of the following: What produces emotions? Why do we have emotions? How do we have emotions? Why do emotional states feel like something?

This book evolved from my earlier book *The Brain and Emotion* (Rolls 1999a) in some of the following ways:

Emotion Explained goes beyond brain mechanisms of emotion, in that it seeks to explain emotions in terms of the following: What produces emotions? (The general answer I propose is reinforcing stimuli, that is rewards and punishers, but with other factors too.) Why do we have emotions? (The overall answer I propose is that emotions are evolutionarily adaptive as they provide an efficient way for genes to influence our behaviour to increase their success.) How do we have emotions? (I answer this by describing what is known about the brain mechanisms of emotion.) Why do emotional states feel like something? This is part of the large problem of consciousness, which I address in Chapter 10. It is in this sense that a broad-ranging explanation of emotion going beyond the brain mechanisms of emotion is the theme of this book.

Emotion Explained goes beyond the brain mechanisms of emotion by developing my approach and theory of the nature of emotion, and comparing my approach to a range of different approaches to the nature of emotion, including the approaches of A.Damasio, J.LeDoux, J.Panksepp, and appraisal theorists such as K.Scherer.

Another way in which this book goes beyond brain mechanisms of emotion is to propose in Chapter 3 a Darwinian account of why animals (including humans) have emotions. The theory will I believe stand the test of time, in the same way as Darwin's theory of evolution by natural selection, and argues that emotions have the important evolutionary role of enabling genes to specify the goals (i.e. the rewards etc that produce emotions) for actions, rather than the actions themselves. The advantage of this Darwinian design is that although the genes specify the goals, the actual actions are not prespecified by the genes, so that there is great flexibility of the actions themselves. This provides a new approach to the nature vs nurture debate in animal behaviour, for it shows how genes can influence behaviour without specifying a fixed, instinctive, behavioural response. I hope that this will make the book of interest to a wide audience, including many interested in evolution and evolutionary biology.

Although in evolution Darwinian processes lead to gene-defined goals, it is also the case that in humans goals may be influenced by other processes, including cultural processes. Indeed, some goals are defined within a culture, for example writing a novel like one by

Tolstoy vs one by Virginia Woolf. But it is argued that it is primary reinforcers specified by genes of the general type shown in Table 2.1 on page 18 that make us want to be recognised in society because of the advantages this can bring, to solve difficult problems, etc, and therefore to perform actions such as writing novels (see further Ridley (2003) Chapter 8, Ridley (1993a) pp. 310 ff, Laland and Brown (2002) pp. 271 ff, and Dawkins (1982)). Indeed, culture is influenced by human genetic propensities, and it follows that human cognitive, affective, and moral capacities are the product of a unique dynamic known as *gene-culture coevolution* (Gintis 2007, Bowles and Gintis 2005, Gintis 2003, Boyd, Gintis, Bowles and Richerson 2003).

We may also note that the theory that genes set many goals for action does not mean that our behaviour is determined by genes. Modern evolutionary theory has led to the understanding that many traits, particularly behavioural ones, may have some genetic basis but that does not mean that they will inevitably appear, because much depends on the environment (Dawkins 1995, Ridley 2003). Further, part of the power of the theory of emotion described here is that in evolution genes specify rewards and punishers that are goals for action, but do not specify the actions themselves, which are flexible and can be learned.

Emotion Explained goes beyond the brain mechanisms of emotion with a treatment (in Chapter 4) of the many different learning processes that become engaged in relation to emotion. The book also includes a formal treatment (in Appendix 1) of reinforcement learning and temporal difference (TD) learning, which are increasingly being used to understand emotion-related learning, as well as its brain mechanisms.

Emotion Explained goes beyond the brain mechanisms of emotion with a treatment of the functions of affective states in motivated behaviour (including hunger, thirst, and sexual behaviour), and indeed proposes a fundamental and simple relation between emotion and motivation. The role of sexual selection in the evolution of affective behaviour is included in Chapter 9.

The book has an integrated section on decision-making (in Chapter 11), and includes links to the developing new field of neuroeconomics.

At the same time, *Emotion Explained* does consider research on how emotion is implemented in the brain, including much new research in the areas of neurophysiology, and functional neuroimaging and clinical neuropsychology in humans. This treatment of the brain mechanisms of emotion is important not only for providing a basis for understanding disorders of emotion, but also turns out to be important in unravelling the many different ways in which emotions can influence our behaviour, because the different brain mechanisms themselves are being unravelled. The book includes a new theory of how the orbitofrontal cortex supports rapid reversals of emotional behaviour, by using a short term memory network for the current rule which acts in a biased competition mode to influence neurons known to be present in the orbitofrontal cortex. This helps to provide a contrast between the functions of the orbitofrontal cortex and amygdala in emotion. A description of the theory is given in Chapter 4, and a formal treatment of how the system operates is given in Appendix 2.

Appendix 2 also shows how it is possible to model the processing involved in emotional learning from the synaptic and neuronal level up through the neuronal network level to predict fMRI neuroimaging signals and behaviour, and thus illustrates a foundation for linking the many different levels of investigation of the brain mechanisms of emotion into a consistent account of precisely how findings at these different levels of exploration are related to each other. This cross-disciplinary approach is a feature of this book. Appendix 1 includes a treat-

ment of autoassociation attractor networks that can maintain stable activity in a brain region, and shows how interacting attractor networks help to provide a foundation for understanding the interactions between mood, and cognition and memory.

The book links to research in psychiatry, with for example discussions of the impulsive behaviour that is a feature of borderline personality disorder, and to research in neurology, with for example assessment of the effects on emotion of damage produced by discrete lesions of the human brain.

Emotion Explained also goes beyond the brain mechanisms involved in emotion, by addressing (in Chapter 10) emotional feelings, part of the much larger problem of consciousness. One issue developed here is the concept that there is a credit assignment problem if a multiple step plan does not succeed, and that higher order thoughts provide a solution to this problem. The book also describes many recent functional neuroimaging investigations in which it has been possible to show that the activations of some brain regions are directly correlated with subjective feelings of affective state.

The material in this text is the copyright of Edmund T. Rolls. Part of the material described in the book reflects work done over many years in collaboration with many colleagues, whose tremendous contributions are warmly appreciated. The contributions of many will be evident from the references cited in the text. In addition, I have benefited enormously from the discussions I have had with a large number of colleagues and friends, many of whom I hope will see areas of the text that they have been able to illuminate. Much of the work described would not have been possible without financial support from a number of sources, particularly the Medical Research Council of the UK, the Human Frontier Science Program, the Wellcome Trust, the McDonnell-Pew Foundation, and the Commission of the European Communities.

The book was typeset in Latex using the WinEdt editor by the author.

The cover shows part of the picture ‘Psyche Opening the Door into Cupid’s Garden’ painted in 1904 by John William Waterhouse.

Updates to the publications cited in this book are available at <http://www.oxcns.org>.

Edmund T. Rolls dedicates this work to the overlapping group: his family, friends, and colleagues: *in salutem praesentium, in memoriam absentium*.

Contents

1	Introduction: the issues	1
1.1	Introduction	1
1.2	Rewards and punishers	2
1.3	The approaches taken to emotion and motivation	5
1.4	The plan of the book	7
2	The nature of emotion	10
2.1	Introduction	10
2.2	A theory of emotion	11
2.3	Different emotions	13
2.4	Refinements of the theory of emotion	21
2.5	The classification of emotion	25
2.6	Other theories of emotion	26
2.6.1	The James–Lange and other bodily theories	26
2.6.2	Appraisal theory	30
2.6.3	Dimensional and categorical theories of emotion	31
2.6.4	Other approaches to emotion	31
2.7	Individual differences in emotion, personality, and emotional intelligence	32
2.8	Cognition and Emotion	35
2.9	Emotion, motivation, reward, and mood	36
2.10	The concept of emotion	37
2.11	Advantages of the approach to emotion described here (Rolls' theory of emotion)	38
3	The functions of emotion:	
	reward, punishment, and emotion in brain design	41
3.1	Introduction	41
3.2	Brain design and the functions of emotion	43
3.2.1	Taxes, rewards, and punishers: gene-specified goals for actions, and the flexibility of actions	43
3.2.2	Explicit systems, language, and reinforcement	47
3.2.3	Special-purpose design by an external agent vs evolution by natural selection	48
3.3	Selection of behaviour: cost–benefit ‘analysis’	49

3.4	Further functions of emotion	51
3.4.1	Autonomic and endocrine responses	51
3.4.2	Flexibility of behavioural responses	52
3.4.3	Emotional states are motivating	53
3.4.4	Communication	54
3.4.5	Social attachment	57
3.4.6	Separate functions for each different primary reinforcer	57
3.4.7	The mood state can influence the cognitive evaluation of moods or memories	58
3.4.8	Facilitation of memory storage	58
3.4.9	Emotional and mood states are persistent, and help to produce persistent motivation	59
3.4.10	Emotions may trigger memory recall and influence cognitive processing	59
3.5	The functions of emotion in an evolutionary, Darwinian, context	59
3.6	The functions of motivation in an evolutionary, Darwinian, context	61
3.7	Are all goals for action gene-specified?	62
4	The brain mechanisms underlying emotion	63
4.1	Introduction	63
4.2	Overview	63
4.3	Representations of primary reinforcers	66
4.3.1	Taste	67
4.3.2	Smell	67
4.3.3	Pleasant and painful touch	67
4.3.4	Visual stimuli	69
4.4	Representing potential secondary reinforcers	71
4.4.1	The requirements of the representation	71
4.4.2	High capacity	74
4.4.3	Objects, and not their reward and punishment associations, are represented in the inferior temporal visual cortex	75
4.4.4	Object representations	77
4.4.5	Invariant representations of faces and objects in the inferior temporal visual cortex	78
4.4.6	Face expression, gesture and view represented in a population of neurons in the cortex in the superior	

temporal sulcus	89
4.4.7 The brain mechanisms that build the appropriate view-invariant representations of objects required for learning emotional responses to objects, including faces	89
4.5 The orbitofrontal cortex	91
4.5.1 Historical background	91
4.5.2 Topology	92
4.5.3 Connections	93
4.5.4 Effects of damage to the orbitofrontal cortex	95
4.5.5 Neurophysiology and functional neuroimaging of the orbitofrontal cortex	97
4.5.6 The human orbitofrontal cortex	131
4.5.7 A neurophysiological and computational basis for stimulus–reinforcer association learning and reversal in the orbitofrontal cortex	140
4.5.8 Executive functions of the orbitofrontal cortex	147
4.6 The amygdala	149
4.6.1 Associative processes involved in emotion-related learning	149
4.6.2 Connections of the amygdala	155
4.6.3 Effects of amygdala lesions	157
4.6.4 Neuronal activity in the primate amygdala to reinforcing stimuli	164
4.6.5 Responses of these amygdala neurons to novel stimuli that are reinforcing	170
4.6.6 Neuronal responses in the amygdala to faces	172
4.6.7 Evidence from humans	175
4.6.8 Amygdala summary	178
4.7 The cingulate cortex	179
4.7.1 Perigenual cingulate cortex and affect	181
4.7.2 Mid-cingulate cortex, the cingulate motor area, and action–outcome learning	185
4.8 Human brain imaging investigations of mood and depression	187
4.9 Output pathways for emotional responses	188
4.9.1 The autonomic and endocrine systems	188
4.9.2 Motor systems for implicit responses, including the basal ganglia	189

4.9.3	Output systems for explicit responses to emotional stimuli	190
4.9.4	Basal forebrain and hypothalamus	191
4.9.5	Basal forebrain cholinergic neurons	191
4.9.6	Noradrenergic neurons	194
4.10	Effects of emotion on cognitive processing and memory	194
4.11	Laterality effects in human emotional processing	200
4.12	Summary	202
4.13	Colour plates	205
5	Hunger	221
5.1	Introduction	221
5.2	Peripheral signals for hunger and satiety	221
5.3	The control signals for hunger and satiety	224
5.3.1	Sensory-specific satiety	224
5.3.2	Gastric distension	230
5.3.3	Duodenal chemosensors	230
5.3.4	Glucostatic hypothesis	230
5.3.5	Body fat regulation – leptin or OB protein	231
5.3.6	Conditioned appetite and satiety	232
5.4	The brain control of eating and reward	233
5.4.1	The hypothalamus	233
5.4.2	Brain mechanisms for the reward produced by the taste of food	243
5.4.3	Convergence between taste and olfactory processing to represent flavour	253
5.4.4	Brain mechanisms for the reward produced by the odour of food	254
5.4.5	The responses of orbitofrontal cortex taste and olfactory neurons to the sight of food	259
5.4.6	Functions of the amygdala and temporal cortex in feeding	259
5.4.7	Functions of the orbitofrontal cortex in feeding	263
5.4.8	Functions of the striatum in feeding	266
5.5	Obesity, bulimia, and anorexia	271
5.6	Conclusions on reward, affective responses to food, and the control of appetite	273
6	Thirst	274
6.1	Introduction	274
6.2	Cellular stimuli for drinking	275

6.3	Extracellular thirst stimuli	276
6.3.1	Extracellular stimuli for thirst	276
6.3.2	Role of the kidney in extracellular thirst: the renin– angiotensin system	278
6.3.3	Cardiac receptors for thirst	279
6.4	Control of normal drinking	279
6.5	Reward and satiety signals for drinking	282
6.6	Summary	286
7	Brain-stimulation reward	288
7.1	Introduction	288
7.2	The nature of the reward produced	288
7.3	The location of brain-stimulation reward sites in the brain	292
7.4	The effects of brain lesions on intracranial self-stimulation	293
7.5	The neurophysiology of reward	294
7.5.1	Lateral hypothalamus and substantia innominata	294
7.5.2	Orbitofrontal cortex	296
7.5.3	Amygdala	298
7.5.4	Nucleus accumbens	299
7.5.5	Central gray of the midbrain	299
7.6	Some of the properties of brain-stimulation reward	300
7.6.1	Lack of satiety with brain-stimulation reward	300
7.6.2	Rapid extinction	302
7.6.3	Priming	302
7.7	Stimulus-bound motivational behaviour	304
7.8	Conclusions	305
7.9	Apostasis	306
8	Pharmacology of emotion, reward, and addiction; the basal ganglia	308
8.1	Introduction	308
8.2	The noradrenergic hypothesis	311
8.3	Dopamine and reward	312
8.3.1	Dopamine and electrical self-stimulation of the brain	312
8.3.2	Self-administration of dopaminergic substances, and addiction	314
8.3.3	Behaviours associated with the release of dopamine	316
8.3.4	The activity of dopaminergic neurons and reward	318
8.4	The basal ganglia	321
8.4.1	Systems-level architecture of the basal ganglia	322
8.4.2	Effects of basal ganglia damage	323

8.4.3	Neuronal activity in the striatum	325
8.4.4	What computations are performed by the basal ganglia?	339
8.4.5	How do the basal ganglia perform their computations?	340
8.4.6	Synthesis on the role of dopamine in reward and addiction	348
8.4.7	Synthesis: emotion, dopamine, reward, punishment, and action selection in the basal ganglia	350
8.5	Opiate reward systems, analgesia, and food reward	352
8.6	Pharmacology of depression in relation to brain systems involved in emotion	353
8.7	Pharmacology of anxiety in relation to brain systems involved in emotion	354
8.8	Cannabinoids	355
8.9	Overview of behavioural selection and output systems involved in emotion	355
9	Sexual behaviour, reward, and brain function; sexual selection of behaviour	358
9.1	Introduction	358
9.2	Mate selection, attractiveness, and love	360
9.2.1	Female preferences	361
9.2.2	Male preferences	363
9.2.3	Pair-bonding, and Love	366
9.3	Parental attachment, care, and parent–offspring conflict	367
9.4	Sperm competition and its consequences for sexual behaviour	368
9.5	Concealed ovulation and its consequences for sexual behaviour	375
9.6	Sexual selection of sexual and non-sexual behaviour	376
9.6.1	Sexual selection and natural selection	376
9.6.2	Non-sexual characteristics may be sexually selected for courtship	379
9.7	Individual differences in sexual rewards	381
9.7.1	Overview	381
9.7.2	How might different types of behaviour be produced by natural selection altering the relative reward value of different stimuli in different individuals?	384
9.7.3	How being tuned to different types of reward could	

help to produce individual differences in sexual behaviour	386
9.8 The neural reward mechanisms that might mediate some aspects of sexual behaviour	387
9.9 Neural basis of sexual behaviour	395
9.10 Conclusion	398
10 Emotional feelings and consciousness: a theory of consciousness	400
10.1 Introduction	400
10.2 A theory of consciousness	401
10.3 Dual routes to action	411
10.4 Content and meaning in representations	418
10.5 Discussion	420
10.6 Conclusions and comparisons	423
11 Conclusions, and broader issues	426
11.1 Conclusions	426
11.2 Decision-making	431
11.2.1 Selection of mainly autonomic responses, and their classical conditioning	431
11.2.2 Selection of approach or withdrawal, and their classical conditioning	431
11.2.3 Selection of fixed stimulus–response habits	432
11.2.4 Selection of arbitrary behaviours to obtain goals, action–outcome learning, and emotional learning	432
11.2.5 The roles of the prefrontal cortex in decision-making and attention	433
11.2.6 Neuroeconomics, reward value, and expected utility	440
11.2.7 Selection of actions by explicit rational thought	444
11.3 Emotion and ethics	445
11.4 Emotion and literature	449
11.5 Close	452
A Neural networks and emotion-related learning	454
A.1 Neurons in the brain, the representation of information, and neuronal learning mechanisms	454
A.1.1 Introduction	454
A.1.2 Neurons in the brain, and their representation in neuronal networks	454

A.1.3	A formalism for approaching the operation of single neurons in a network	456
A.1.4	Synaptic modification	458
A.1.5	Long-Term Potentiation and Long-Term Depression	459
A.1.6	Distributed representations	464
A.2	Pattern association memory	466
A.2.1	Architecture and operation	466
A.2.2	A simple model	469
A.2.3	The vector interpretation	472
A.2.4	Properties	472
A.2.5	Prototype extraction, extraction of central tendency, and noise reduction	475
A.2.6	Speed	475
A.2.7	Local learning rule	476
A.2.8	Implications of different types of coding for storage in pattern associators	482
A.3	Autoassociation memory: attractor networks	483
A.3.1	Architecture and operation	483
A.3.2	Introduction to the analysis of the operation of autoassociation networks	485
A.3.3	Properties	486
A.4	Coupled attractor networks	491
A.5	Reinforcement learning	493
A.5.1	Associative reward–penalty algorithm of Barto and Sutton	494
A.5.2	Error correction or delta rule learning, and classical conditioning	496
A.5.3	Temporal Difference (TD) learning	497
B	Reward reversal in the orbitofrontal cortex – a model	501
B.1	Introduction	501
B.2	The model of stimulus–reinforcer association reversal	503
B.2.1	The network	504
B.2.2	Reward reversal: the operation of the rule module neurons	507
B.2.3	The neurons in the model	509
B.2.4	The synapses in the model	510
B.3	Operation of the reward reversal model	511
B.4	A model of reversal of a conditional object-response task by the dorsolateral prefrontal cortex	515

B.5	Evaluation of the models	517
B.6	Integrate-and-Fire model equations and parameters	521
B.7	Simulation of fMRI signals: haemodynamic convolution of synaptic activity	522
C	Glossary	525
	References	528
	Index	602

2 The nature of emotion

2.1 Introduction

What are emotions? This is a question in which almost everyone is interested. There have been many answers, many of them surprisingly unclear and ill-defined. William James (1884) was at least clear about what he thought. He believed that emotional experiences were produced by sensing bodily changes, such as changes in heart rate or in skeletal muscles (the muscles involved in voluntary movements). His view was that “We feel frightened because we are running away”. But he left unanswered the crucial question even for his theory, which is: Why do some events make us run away (and then feel emotional), whereas others do not?

A more modern theory is that of Frijda (1986), who argues that a change in action readiness is the central core of an emotion. Oatley and Jenkins (1996) (page 96) make this part of their definition too, stating that “the core of an emotion is readiness to act and the prompting of plans”. But surely subjects in reaction time experiments in psychology who are continually for thousands of trials altering their action readiness are very far indeed from having normal or strong emotional experiences? Similarly, we can perform an action in response to a verbal request (e.g. open a door), yet may not experience great emotion when performing this action. Another example might be the actions that are performed in driving a car on a routine trip – we get ready, and many actions are performed, often quite automatically, yet little emotion occurs. So it appears that there is no necessary link between performing actions and emotion. This may not be a clear way to define emotion.

Because it is important to be able to specify what emotions are, in this Chapter we consider a systematic approach to this question. Part of the approach is to ask what causes emotions. Can clear conditions be specified for the circumstances in which emotions occur? This is considered in Section 2.2. Continuing with this theme, when we have come to understand the conditions under which emotions occur, does this help us to classify and describe different emotions systematically, in terms of differences between the different conditions that cause emotions to occur. A way in which a systematic account of different emotions can be provided is described in Section 2.3. A major help in understanding emotions would be provided by understanding what the functions of emotion are. It turns out that emotions have quite a number of different functions, each of which helps us to understand emotions a little more clearly. These different functions of emotion are described in Chapter 3. Understanding the different functions of emotion helps us to understand also the brain mechanisms of emotion, for it helps us to see that emotion can operate to affect several different output systems of the brain.

These analyses leave open though a major related question, which is why emotional states feel like something to us. This it transpires is part of the much larger, though more speculative, issue of consciousness, and why anything should feel like something to us. This aspect of emotional feelings, because it is part of the much larger issue of consciousness, is deferred until Chapter 10.

In Chapter 2, in considering the function of emotions, the idea is presented that emotions are part of a system that helps to map certain classes of stimuli, broadly identified as rewarding and punishing stimuli (i.e. aversive stimuli or ‘punishers’), to action systems. Part of the idea is that this enables a simple interface between such stimuli and actions. This is an important area in its own right, which goes to the heart of why animals are built to respond to rewards and punishments, and have emotions.

The suggestion made in this book is that we now have a way of systematically approaching the nature of emotions, their functions, and their brain mechanisms. Doubtless in time there will be changes and additions to the overall picture. But the suggestion is that the ideas and theory presented here do provide a firm and systematic foundation for understanding emotions, their functions, and their brain mechanisms in a well-founded evolutionary context.

2.2 The outline of a theory of emotion

I will first introduce the essence of the definition of emotion that I propose. *The definition of emotions is that emotions are states elicited by rewards and punishers, that is, by instrumental reinforcers.* As described in Section 1.2, a reward is anything for which an animal will work. A punisher is anything that an animal will work to escape or avoid, or that will suppress actions on which it is contingent³. I note that any change in the regular delivery of a reward or a punisher acts as a reinforcer. The relevant states elicited by the reinforcers are those with the particular functions described in Chapter 3.

An example of an emotion might thus be happiness produced by being given a reward, such as a hug, a pleasant touch, praise, winning a large sum of money, or being with someone whom one loves. All these things are rewards, in that we will work to obtain them. Another example of an emotion might be fear produced by the sound of a rapidly approaching bus when we are cycling, or the sight of an angry expression on someone’s face. We will work to avoid such stimuli, which are punishers. Another example might be frustration, anger, or sadness produced by the omission of an expected reward such as a prize, or the termination of a reward such as the death of a loved one. Another example might be relief, produced by the omission or termination of a punishing stimulus, for example the removal of a painful stimulus, or sailing out of danger. These examples indicate how emotions can be produced by the delivery, omission, or termination of rewarding or punishing stimuli, and go some way to indicate how different emotions could be produced and classified in terms of the rewards and punishers received, omitted, or terminated.

Before accepting this proposal, we should consider whether there are any exceptions to the proposed rule. Indeed, at first this may appear to be a rather reductionist hypothesis about what produces emotions. However, one way to test the suggested definition of the events that cause emotions is to ask whether there are any rewards or punishers that do not produce emotions. Conversely, we should ask whether there are any emotions that are produced by stimuli, events, or remembered events that are not rewarding or punishing. If we cannot find exceptions, then we should accept the suggestion as a useful identification, summary, and working definition of the conditions that produce emotions. Therefore in the next few pages we consider the questions: ‘Are any emotions caused by stimuli, events, or remembered events that are not rewarding or punishing? Do any rewarding or punishing stimuli not cause

³A full definition in terms of reinforcement contingencies is given below.

emotions?’ But first it is worth pointing out that in fact many approaches to or theories of emotion have in common that part of the process involves ‘appraisal’ (e.g. Frijda (1986); Oatley and Johnson-Laird (1987); Lazarus (1991); Izard (1993); Stein, Trabasso and Liwag (1994)). This is part, for example, of the suggestion made by Oatley and Jenkins (1996), who on page 96 write that “an emotion is usually caused by a person consciously or unconsciously evaluating an event as relevant to a concern (a goal) that is important; the emotion is felt as positive when a concern is advanced and negative when a concern is impeded”. The concept of appraisal presumably involves in all these theories assessment of whether something is rewarding or punishing, that is whether it will be worked for or avoided. The description in terms of reward or punisher adopted here simply seems much more precisely and operationally specified.

The idea that rewards and punishers, that is instrumental reinforcers, are the stimuli that produce emotions has a considerable history, with origins that can be traced back to Watson (1929), Watson (1930), Harlow and Stagner (1933), Amsel (1958), and Amsel (1962). More recently, the approach was developed by Millenson (1967), Weiskrantz (1968), and Jeffrey Gray (1975, 1981). We can introduce some of the emotions that result from different reinforcement contingencies as follows. Consider the emotional effects of delivery of a ‘reward’ : a state such as pleasure or happiness will be produced. An example might be receiving a prize for excellent work. Now consider the emotional effects of delivery of a ‘punisher’ : pain or fear may be produced. For example, fear is an emotional state that might be produced by a sound that has previously been associated with a painful electrical shock. Shock in this example is the primary reinforcer, and fear is the emotional state that occurs to the tone stimulus as a result of the learning of the stimulus (i.e. tone)–reinforcer (i.e. shock) association. The tone in this example is a conditioned stimulus because of stimulus–reinforcer association learning, and has secondary reinforcing properties in that responses will be made to escape from it and thus avoid the primary reinforcer, shock.

The converse reinforcement contingencies produce the opposite effects on behaviour, and produce different emotions. The omission or termination of a reward (‘extinction’ and ‘time out’ respectively) reduce the probability of responses, and may produce the emotions of frustration, disappointment, or rage. (Imagine not receiving a prize that you deserved.) Behavioural responses followed by the omission or termination of a punisher increase in probability (this pair of reinforcement operations being termed ‘active avoidance’ and ‘escape’, respectively), and are associated with emotions such as relief.

The classification of emotions in terms of reinforcement contingencies is developed further in Section 2.3, and more formal definitions of rewards and punishers, and how they are related to learning theory concepts such as reinforcement and punishment are given in the footnote ⁴,

⁴Instrumental reinforcers are stimuli that, if their occurrence, termination, or omission is made contingent upon the making of an action, alter the probability of the future emission of that action (Gray 1975, Mackintosh 1983, Dickinson 1980, Lieberman 2000). Rewards and punishers are instrumental reinforcing stimuli. The notion of an action here is that an arbitrary action, e.g. turning right vs turning left, will be performed in order to obtain the reward or avoid the punisher, so that there is no pre-wired connection between the response and the reinforcer. Some stimuli are primary (unlearned) reinforcers (e.g., the taste of food if the animal is hungry, or pain); while others may become reinforcing by learning, because of their association with such primary reinforcers, thereby becoming ‘secondary reinforcers’. This type of learning may thus be called ‘stimulus–reinforcer association’, and occurs via an associative learning process. A positive reinforcer (such as food) increases the probability of emission of a response on which it is contingent, the process is termed **positive reinforcement**, and the outcome is a reward (such as food). A negative reinforcer (such as a painful stimulus) increases the probability of emission of a response that causes the negative reinforcer to be omitted (as in active avoidance) or terminated (as in escape), and the procedure is termed **negative**

and in Sections 1.2 and 4.6.1. My argument is that an affectively positive or ‘appetitive’ stimulus (which produces a state of pleasure) acts operationally as a **reward**, which when delivered acts instrumentally as a positive reinforcer, or when not delivered (omitted or terminated) acts to decrease the probability of responses on which it is contingent. Conversely I argue that an affectively negative or aversive stimulus (which produces an unpleasant state) acts operationally as a **punisher**, which when delivered acts instrumentally to decrease the probability of responses on which it is contingent, or when not delivered (escaped from or avoided) acts as a negative reinforcer in that it then increases the probability of the action on which its non-delivery is contingent⁵.

The link between emotion and instrumental reinforcers being made is partly an operational link. Most people find that it is not easy to think of exceptions to the statements that emotions occur after rewards or punishers are given (sometimes continuing for long after the eliciting stimulus has ended, as in a mood state); or that rewards and punishers, but not other stimuli, produce emotional states. But the link is deeper than this, as we will see, in that the theory has been developed that genes specify primary reinforcers in order to encourage the animal to perform arbitrary actions to seek particular goals, thus increasing the probability of their own (the genes’) survival into the next generation (Rolls 1999a). The emotional states elicited by the reinforcers have a number of functions, described below, related to these processes.

Before considering how different emotions are related to different reinforcement contingencies in Section 2.3, I clarify a matter of terminology about moods vs emotions. A useful convention to distinguish between emotion and a mood state is as follows. An emotion consists of cognitive processing that results in a decoded signal that an environmental event (or remembered event) is reinforcing, together with the mood state produced as a result. If the mood state is produced in the absence of the external sensory input and the cognitive decoding (for example by direct electrical stimulation of the brain, see Chapter 7), then this is described only as a mood state, and is different from an emotion in that there is no object in the environment towards which the mood state is directed. (In that emotions are produced by stimuli or objects, and thus emotions ‘take or have an object’, emotional states are examples of what philosophers call intentional states.) It is useful to emphasize that there is great opportunity for cognitive processing (whether conscious or not) in emotions, for cognitive processes will very often be required to determine whether an environmental stimulus or event is reinforcing (see further Section 2.4).

2.3 Different emotions

As introduced in Section 2.2, the different emotions can in part be described and classified according to whether the reinforcer is positive or negative, and by the reinforcement contingency. An outline of such a classification scheme, elaborated by Rolls (1990d), Rolls (1999a)

reinforcement. In contrast, **punishment** refers to procedures in which the probability of an action is decreased. Punishment thus describes procedures in which an action decreases in probability if it is followed by a painful stimulus, as in passive avoidance. Punishment can also be used to refer to a procedure involving the omission or termination of a reward (‘extinction’ and ‘time out’ respectively), both of which decrease the probability of responses (Gray 1975, Mackintosh 1983, Dickinson 1980, Lieberman 2000).

⁵Note that my definition of a punisher, which is similar to that of an aversive stimulus, is of a stimulus or event that can either decrease the probability of actions on which it is contingent, or increase the probability of actions on which its non-delivery is contingent. The term punishment is restricted to situations where the probability of an action is being decreased.

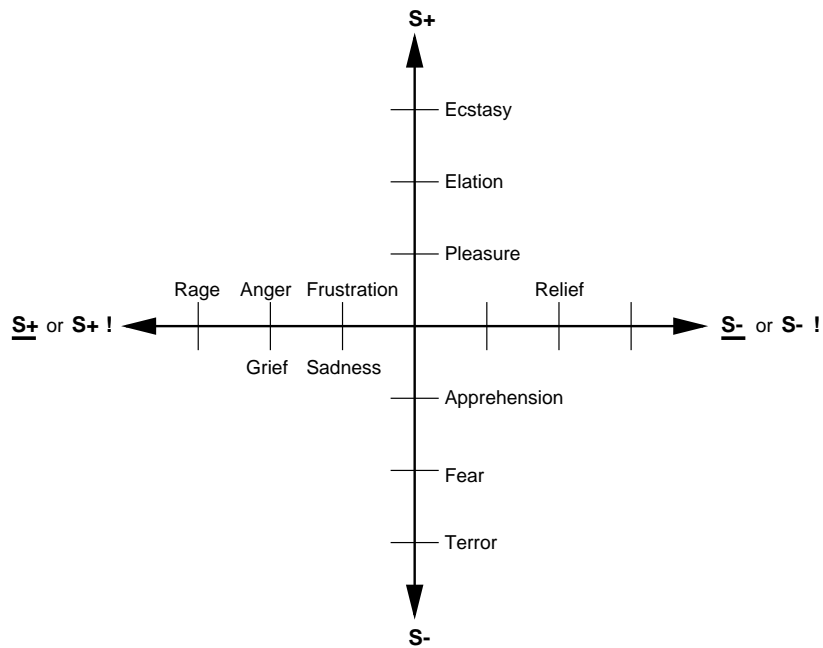


Fig. 2.1 Some of the emotions associated with different reinforcement contingencies are indicated. Intensity increases away from the centre of the diagram, on a continuous scale. The classification scheme created by the different reinforcement contingencies consists with respect to the action of (1) the delivery of a reward (S+), (2) the delivery of a punisher (S-), (3) the omission of a reward ($\underline{S+}$) (extinction) or the termination of a reward (S+!) (time out), and (4) the omission of a punisher ($\underline{S-}$) (avoidance) or the termination of a punisher (S-!) (escape). Note that the vertical axis describes emotions associated with the delivery of a reward (up) or punisher (down). The horizontal axis describes emotions associated with the non-delivery of an expected reward (left) or the non-delivery of an expected punisher (right).

and Rolls (2000f), is shown in Fig. 2.1. Movement away from the centre of the diagram represents increasing intensity of emotion, on a continuous scale. The diagram shows that emotions associated with the delivery of a reward (S+) include pleasure, elation and ecstasy. Of course, other emotional labels can be included along the same axis. Emotions associated with the delivery of a punisher (S-) include apprehension, fear, and terror (see Fig. 2.1). Emotions associated with the omission of a reward ($\underline{S+}$) or the termination of a reward (S+!) include frustration, anger and rage. Emotions associated with the omission of a punisher ($\underline{S-}$) or the termination of a punisher (S-!) include relief. Although the classification of emotions presented here (and by Rolls (1986c), Rolls (1986a), Rolls (1990d) and Rolls (1999a)) differs from earlier theories, the approach adopted here of defining and classifying emotions by reinforcing effects is one that has been developed in a number of earlier analyses (e.g. Millenson (1967), Gray (1975), Gray (1981); see Strongman (2003)).

I should make it clear that the scheme shown in Fig. 2.1 is not intended to be a dimensional scheme. [A dimensional scheme is one in which independent factors or dimensions have been identified that account for the major and independent sources of variation in a data set. Some investigators then work to show that these dimensions can be interpreted both biologically (for example as differing in autonomic, endocrine, or arousal-related ways) and psychologically

(e.g. as representing anger vs fear), as described in Section 2.6.3.] However, the import of what is shown in Fig. 2.1 is to set out a set of logical possibilities of ways in which reinforcement contingencies can vary, and to show how they may be related to some different types of emotion.

It is actually a possibility that the four directions shown in Fig. 2.1 are at least partly independent from each other, and that a four-dimensional space is spanned by what is shown in Fig. 2.1. For example, sensitivity to (that is the ability to respond to) reward (S+) could be at least partly independent from sensitivity to punishers (S-), sensitivity to non-reward (\underline{S}_+ and S+!), and sensitivity to non-delivery of a punisher (\underline{S}_- and S-!). The dimensions or independent ways in which emotions may differ from each other could thus span 4 dimensions even with what is shown in Fig. 2.1, and these ways are expanded greatly as shown by the following further effects that make different emotions different to each other.

One important point about Fig. 2.1 is that there are a large number of different primary reinforcers, and that for example the reward label S+ shows states that might be elicited by just one type of reward, such as a pleasant touch. There will be a different reward axis (S+) and non-reward axis (\underline{S}_+ and S+!) for each type of reward (e.g. pleasant touch vs sweet taste); and, correspondingly, a different punisher axis (S-) and non-punisher axis (\underline{S}_- and S-!) for each type of punisher (e.g. pain vs bitter taste).

Different reinforcement contingencies can thus be used to classify a wide range of emotions. However, some of my tutorial pupils at Oxford sometimes expressed the view that reinforcement contingencies alone might not be able to account for the full range of human emotions. I therefore set out for them ways in which a system based on reinforcement contingencies could be developed in a number of different ways to give an account of most emotions. This extended set of ways of accounting for different emotions was published in 1986 (Rolls (1986c), Rolls (1986a)), and developed a little in later publications (e.g. Rolls (1995b), Rolls (1999a)). It is described, and elaborated further next. If the reader can think of any emotions that cannot be accounted for by a combination of the ways described next, then it would be interesting to consider what further extensions might be needed.

1. Reinforcement contingency

The first way in which different classes of emotion could arise is because of different reinforcement contingencies, as described above and indicated in Fig. 2.1.

2. Intensity

Second, different intensities within these classes can produce different degrees of emotion (see above and Millenson (1967)). For example, as the strength of a positive reinforcer being presented increases, emotions might be labelled as pleasure, elation, and ecstasy. Similarly, as the strength of a negative reinforcer being presented increases, emotions might be labelled as apprehension, fear, and terror (see Fig. 2.1). It may be noted here that anxiety can refer to the state produced by stimuli associated with the non-delivery of a reward or the delivery of a punisher (Gray 1987).

3. Multiple reinforcement associations

Third, any environmental stimulus might have a number of different reinforcement associations. For example, a stimulus might be associated both with the presentation of a reward and of a punisher, allowing states such as conflict and guilt to arise. The different possible

combinations greatly increase the number of possible emotions.

4. Different primary reinforcers

Fourth, emotions elicited by stimuli associated with different primary reinforcers will be different even within a reinforcement category (i.e. with the same reinforcement contingency), because the original reinforcers are different. Thus, for example, the state elicited by a stimulus associated with a reward such as the taste of food will be different from that elicited by a reward such as being groomed. Indeed, it is an important feature of the association memory mechanisms described here that when a stimulus is applied, it acts as a key which ‘looks up’ or recalls the original primary reinforcer with which it was associated. Thus emotional stimuli will differ from each other in terms of the original primary reinforcers with which they were associated.

A summary of many different primary reinforcers is provided in Table 2.1, and inspection of this will help to show how some different emotions are produced by different primary reinforcers. For example, from Table 2.1 it might be surmised that one of the biological origins of the emotion of jealousy might be the state elicited in a male when his partner is courted by another male, because this threatens his parental investment in the offspring he raises with his partner, as described in Chapter 9. Jealousy in females would arise in a corresponding way. Examples of how further emotions including guilt, shame, anger, forgiveness, envy and love may arise in relation to particular primary reinforcers are provided later in this Section, throughout this Chapter, in Chapters 3 and 9, and in many other places in this book.

5. Different secondary reinforcers

A fifth way in which emotions can be different from each other is in terms of the particular (conditioned) stimulus that elicits the emotion, and the situation in which it occurs. Thus, even though the reinforcement contingency and even the unconditioned reinforcer may be identical, emotions will still be different cognitively, if the conditioned stimuli that give rise to the emotions are different (that is, if the objects of the emotion are different). For example, the emotional state elicited by the sight of one person may be different from that elicited by the sight of another person because the people, and thus the cognitive evaluations associated with the perception of the stimuli, are different. In another example, not obtaining a monetary reward in a gambling task might lead to frustration, but being blocked by another person from obtaining a reward might lead to anger directed at the person.

Thus evolution may have shaped different reinforcers to contribute in different ways and depending on the environmental circumstances to the exact emotion produced. For example, some emotions may be related to social reinforcers (e.g. love, anger, envy, and breaking rules of society so that shame is produced, see further Section 11.3), others to non-social reinforcers (such as fear of a painful stimulus), and others to solving difficult problems. By taking into account the nature of the primary reinforcer, the nature of the secondary reinforcer, and the environmental circumstances in which these apply, many different emotions can thus be accounted for, and cognitive factors taken into account. The common underlying basis of emotion remains however that it is related to goals/instrumental reinforcers, and the reinforcement contingencies that operate. The variety of different goals, and the contingencies and environmental situations in which they occur, combine to contribute to the richness in the variety of emotional states.

The gene-specified reinforcer approach to emotion advocated in this book is somewhat

different to the domain-specific (vs domain general) approach of some evolutionary biologists (see Nesse (2000b)). In the domain-specific approach, a modular approach to different emotions may be taken, and the temptation is to end with a large number of specialized emotional systems, each promoting particular types of action. In contrast, in the approach described here, different genes build different reinforcement systems that define the goals for actions, and arbitrary actions appropriate for reaching the goal (i.e. instrumental actions) are then performed, with action–outcome learning guiding the actions produced. This can result in a rich variety of actions being selected in different emotion-provoking situations, without a tendency to suggest that particular perhaps instinctive actions are coupled to particular emotions. Instead, ‘instinct’ is involved in the process whereby the *goals* for actions, which are reinforcing stimuli, are specified by genes as a result of natural selection, and the behavioural response itself is not specified or ‘determined’ (see further Section 3.5).

Further, in the approach described here, modular neural systems useful for face identification, face expression recognition, and head gesture and movement may evolve because of the different specialized computational requirements for each and the importance of minimizing wiring length in the brain (see Section 4.4), and because the presence of these systems helps to provide representations that are useful in defining which stimulus or object-related events in the environment are associated with primary reinforcers.

6. The behavioural responses that are available

A sixth possible way in which emotions can vary arises when the environment constrains the types of behavioural response that can be made. For example, if an active behavioural response can occur to the omission of an expected reward, then anger might be produced and directed at the person who prevented the reward being obtained, but if only passive behaviour is possible, then sadness, depression or grief might occur.

By realizing that these six possibilities can occur in different combinations, it can be seen that it is possible to account for a very wide range of emotions, and this is believed to be one of the strengths of the approach described here. It is also the case that the extent to which a stimulus is reinforcing on a particular occasion (and thus an emotion is produced) depends on the prior history of reinforcements (both recently through processes that include sensory-specific satiety, and in the longer term), and that the current mood state can affect the degree to which a stimulus (a term that includes cognitively decoded events and remembered events) is reinforcing (see Section 4.10).

If we wish to consider the number of independent ways in which emotions may differ from each other (for comparison with the ‘dimensional’ theories described in Section 2.6.3) we see immediately that a vast subtlety of emotions can be systematically described using the approach described here. For example, based on the four different reinforcement contingencies shown in Fig. 2.1 we have four at least potentially independent ‘dimensions’, which are combined with perhaps another 100–500 independently varying (in that they are gene-specified) primary reinforcers, some of which are included in Table 2.1. These are combined with constraints to the actions that may be possible when a reinforcer is received (the ‘coping potential’ of appraisal theorists), which potentially at least doubles the number of emotions that can be described. We add further combinatorial possibilities by noting (point 3 above) that a given stimulus in the world may have many different reinforcement associations producing states such as conflict. The possible number of different emotions can be further multiplied by

Table 2.1 Some primary reinforcers, and the dimensions of the environment to which they are tuned**Taste**

Salt taste	reward in salt deficiency
Sweet	reward in energy deficiency
Bitter	punisher, indicator of possible poison
Sour	punisher
Umami	reward, indicator of protein; produced by monosodium glutamate and inosine monophosphate
Tannic acid	punisher; it prevents absorption of protein; found in old leaves; probably somatosensory rather than strictly gustatory; see Critchley and Rolls 1996c

Odour

Putrefying odour	punisher; hazard to health
Pheromones	reward (depending on hormonal state)

Somatosensory

Pain	punisher
Touch	reward
Grooming	reward; to give grooming may also be a primary reinforcer
Washing	reward
Temperature	reward if tends to help maintain normal body temperature; otherwise punisher

Visual

Snakes, etc.	punisher for, e.g., primates
Youthfulness	reward, associated with mate choice
Beauty, e.g. symmetry	reward
Secondary sexual characteristics	rewards
Face expression	reward (e.g. smile) or punisher (e.g. threat)
Blue sky, cover, open space	reward, indicator of safety
Flowers	reward (indicator of fruit later in the season?)

Auditory

Warning call	punisher
Aggressive vocalization	punisher
Soothing vocalization	reward (part of the evolutionary history of music, which at least in its origins taps into the channels used for the communication of emotions)

Table 2.1 continued **Some primary reinforcers, and the dimensions of the environment to which they are tuned****Reproduction**

courtship	reward
sexual behaviour	reward (a number of different reinforcers, including a low waist-to-hip ratio, and attractiveness influenced by symmetry and being found attractive by members of the other sex, are discussed in Chapter 9)
mate guarding	reward for a male to protect his parental investment; jealousy results if his mate is courted by another male, because this may ruin his parental investment
nest building	reward (when expecting young)
parental attachment	reward
infant attachment to parents	reward
crying of infant	punisher to parents; produced to promote successful development

Other

Novel stimuli	rewards (encourage animals to investigate the full possibilities of the multidimensional space in which their genes are operating)
Sleep	reward; minimizes nutritional requirements and protects from danger
Altruism to genetically related individuals	reward (kin altruism)
Altruism to other individuals	reward while the altruism is reciprocated in a 'tit-for-tat' reciprocation (reciprocal altruism) Forgiveness, honesty, and altruistic punishment are some associated heuristics (May provide underpinning for some aspects of what is felt to be moral)
Altruism to other individuals	punisher when the altruism is not reciprocated
Group acceptance, reputation	reward (social greeting might indicate this) These goals can account for why some culturally specified goals are pursued
Control over actions	reward
Play	reward
Danger, stimulation, excitement	reward if not too extreme (adaptive because of practice?)
Exercise	reward (keeps the body fit for action)
Mind reading	reward; practice in reading others' minds, which might be adaptive
Solving an intellectual problem	reward (practice in which might be adaptive)
Storing, collecting	reward (e.g. food)
Habitat preference, home, territory	reward
Some responses	reward (e.g. pecking in chickens, pigeons; adaptive because it is a simple way in which eating grain can be programmed for a relatively fixed type of environmental stimulus)
Breathing	reward

the fact that each primary reinforcer may have associated with it almost any neutral stimulus to produce a secondary reinforcer.

The resulting number of emotional states that can be described and categorized is clearly enormous, even if we do not assume that each of the above factors operates strictly inde-

pendently (factorially). For example, it is likely that if a gene were to specify a particular reward as being particularly intense in an individual, for example the pleasantness of touch, then omitting (S+) or terminating (S+!) this reward might also be expected to be particularly intense, so the contributions of reinforcement contingency and identity of the primary reinforcer might combine additively rather than multiplicatively. Even if there is only partial independence of the different processes 1–6 above, and of variation within each process, then nevertheless many different emotions can be systematically classified and described. It does of course remain an interesting issue of how the processes described above do combine, and of the extent to which a few factors actually do account for a great deal in the variation between different emotions. For example, if in an individual's sensitivity to non-reward is generally much more intense than the individual's sensitivity to reward, then this will shape the emotions in that individual, and account for quite a deal of the variance between that individual's emotional states. Such a factor might also account for quite an amount of the variation in emotions and personality between individuals (see Section 2.7).

Some examples of how different emotions might be classified using the above criteria now follow. Fear is a state that might be produced by a stimulus that has become a secondary reinforcer by virtue of its learned association with a primary negative reinforcer such as pain (see Fig. 2.1). Anger is a state that might be produced by the omission of an expected reward, frustrative non-reward, when an active behavioural response is possible (see Fig. 2.1). (In particular, anger may occur if another individual prevents an expected reward from being obtained.) Guilt may arise when there is a conflict between an available reward and a rule or law of society. Jealousy is an emotion that might be aroused in a male if the faithfulness of his partner seems to be threatened by her liaison (e.g. flirting) with another male. In this case the reinforcement contingency that is operating is produced by a punisher, and it may be that males are specified genetically to find this punishing because it indicates a potential threat to their paternity and paternal investment, as described in Chapters 9 and 3. Similarly, a female may become jealous if her partner has a liaison with another female, because the resources available to the 'wife' useful to bring up her children are threatened. Again, the punisher here may be gene-specified, as described in Chapter 3. Envy or disappointment might be produced if a prize is obtained by a competitor. In this case, part of the way in which the frustrative non-reward is produced is by the cognitive understanding that this is a competition in which there will be a winner, and that the person has set himself or herself the goal of obtaining it.

The partial list of primary reinforcers provided in Table 2.1 should provide readers with a foundation for starting to understand the rich classification scheme for different types of emotion that can be classified in this way.

Many other similar examples can be surmised from the area of evolutionary psychology (see e.g. Ridley (1993b), Buss (1999) and Barrett, Dunbar and Lycett (2002)). For example, there may be a set of reinforcers that are genetically specified to help promote social cooperation and even reciprocal altruism. Such genes might specify that emotion should be elicited, and behavioural changes should occur, if a cooperating partner defects or 'cheats' (Cosmides and Tooby 1999). Moreover, the genes may build brains with genetically specified rules that are useful heuristics for social cooperation, such as acting with a strategy of 'generous tit-for-tat', which can be more adaptive than strict 'tit-for-tat', in that being generous occasionally is a good strategy to help promote further cooperation that has failed when both partners defect in a strict 'tit-for-tat' scenario (Ridley 1996). Genes that specify good heuristics to promote social cooperation may thus underlie such complex emotional states as feeling forgiving.

It is suggested that many apparently complex emotional states have their origins in designing animals to perform well in such sociobiological and socioeconomic situations (Ridley 1996, Glimcher 2003, Glimcher 2004). Indeed, many principles that humans accept as ethical may be closely related to strategies that are useful heuristics for promoting social cooperation, and emotional feelings associated with ethical behaviour may be at least partly related to the adaptive value of such gene-specified strategies. These ideas are developed in Section 11.3.

These examples indicate that an emotional state can be systematically specified and classified using the six principles described above in this Section. The similarity between particular emotions will depend on how close they are in the space defined by the above principles.

2.4 Refinements of the theory of emotion

The definition of emotions given above, that they are states produced by instrumental reinforcing stimuli, and have particular functions, is refined now.

First, when positively reinforcing (rewarding) stimuli (such as the taste of food or water) are relevant to a drive state produced by a change in the *internal milieu* (such as hunger and thirst), then we do not normally classify these stimuli as emotional, though they do produce pleasure, and indeed we describe the state they produce as affective (see Chapters 5 and 6). In contrast, emotional states are normally initiated by reinforcing stimuli that have their origin in the external environment, such as an (external) noise associated with pain (delivered by an external stimulus). We may then have identified a class of reinforcers (in our example, food) that we do not want to say cause emotions. This then is a refinement of the definition of emotions given above. Fortunately, we can encapsulate the set of reinforcing stimuli that we wish to exclude from our definition of stimuli that produce emotion. They are the set of external reinforcers (such as the sight of food) that are relevant to motivational states such as hunger and thirst, which are controlled by internal need-related (i.e. homeostatic) signals such as the concentration of glucose in the plasma (see Chapter 5). However, there is room for plenty of further discussion and refinement here. Perhaps some people (especially French people?) might say that they do experience emotion when they savour a wonderful food. There may well be cultural differences here in the semantics of whether such reinforcing stimuli should be included within the category that produce emotions.

Another area for discussion is how we wish to categorize the reinforcers associated with sexual behaviour. Such stimuli may be made to be rewarding, and to feel pleasurable, partly because of the internal hormonal state. Does this mean that we wish to exclude such stimuli from the class that we call emotion-provoking, in the same way that we might exclude food reward from the class of stimuli that are said to cause emotion, because the reward value of food depends on an internal controlling signal? I am not sure that there is a perfectly clear answer to this. But this may not matter, as long as we understand that there are some rewarding stimuli that some may wish to exclude from those that cause emotional states.

Second, emotional states can be produced by *remembered reinforcing stimuli*. (Indeed, when we remember stimuli or events, many of the cortical areas activated by the original sensory stimulus are also activated by the remembered stimuli or events. This is the case

for most of the cortical areas in each sensory system, apart perhaps from the first (see Rolls (1989a), and Rolls and Treves (1998)). Thus if we recall a particular event, and this leads to reinstatement of activity in the higher parts of the visual system, this activity will provide inputs to the later parts of the brain involved in emotion, so that emotional states may then be produced.

Third, the stimulus that produces the emotional state does not have to be shown to be a reinforcer when producing the emotional state – it simply has to be *capable of being shown to have instrumental reinforcing properties*. An emotion-provoking stimulus can act as a reward or punisher, and is a goal for possible action.

Fourth, the definition given provides great opportunity for *cognitive processing* (whether conscious or not) in emotions, for cognitive processes will very often be required to determine whether an environmental stimulus or event is a reward or punisher. Normally an emotion consists of this cognitive processing that results in a decoded signal that the environmental event is reinforcing, together with the mood state produced as a result. If the mood state is produced in the absence of the external sensory input and the cognitive decoding (for example by direct electrical stimulation of the amygdala, see Rolls (1975) and Rolls (1999a)), then this is described only as a mood state, and is different from an emotion in that there is no object in the environment towards which the mood state is directed. The external reinforcing stimulus may alter the mood state very rapidly, and then the firing of the neurons that represent the mood state may gradually return back to their baseline firing rate, depending on the time course of the emotional state that is produced by the external reinforcing stimulus.

While discussing *mood*, it is worth pointing out that mood may be a particularly difficult state for the brain to maintain at a relatively constant level. In sensory systems the situation is different, for most sensory systems work by contrast, rather than absolute level. For example, early in the visual system it is the difference in brightness levels present at an edge, rather than the absolute brightness that is signaled. This is achieved by a process of lateral inhibition, which means that neighbouring neurons effectively inhibit each other. The result is that it is only at a dark–light boundary, where there is contrast, that neurons are firing. (In fact the firing will be fast on the bright side of the edge, and low, below a spontaneous level of firing, on the dark side of the edge. In the middle of a large bright area few neurons will be active, because the nearby neurons will be inhibiting each other.) However, for mood, the situation may be different. Here, the absolute firing rates of the neurons that represent mood state must be set to fire at the appropriate rate for long periods. Any drift in firing rates would represent a change of mood level. The situation for a brain system that represents mood may thus be different from that involved in most sensory and motor processing in the brain, both because in sensory systems it is the local contrast of firing rate that is important, not the absolute level, and because in sensory systems the inputs keep changing, so that it is not necessary to maintain an absolute value for long. The difficulty of maintaining a constant absolute level of firing in neurons may contribute to ‘spontaneous’ mood swings, depression that occurs without a clear external cause, and the multiplicity of hormonal and transmitter systems that seem to be involved in the control of mood (see Chapters 4 and 8).

Having said this, it also seems to be the case that there is some ‘*regression to a constant value*’ for emotional states. What I mean by this is that we are sensitive to some extent not just to the absolute level of reinforcers being received, but also to the change in the rate, prob-

Positive and Negative Contrast

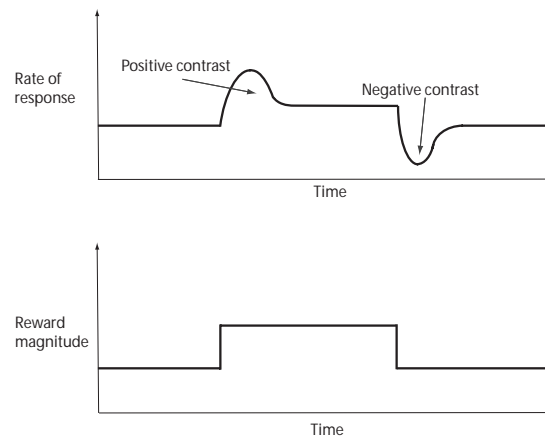


Fig. 2.2 Positive and negative contrast. If the magnitude of a moderate reward (positive reinforcer) is increased, then the rate of working increases markedly, but then drops back to a rate just greater than that to the moderate reinforcer. The positive overshoot is positive contrast. The converse happens if the magnitude of the reward is decreased.

ability, or magnitude of reinforcers being received. This is well shown by the phenomenon of *positive and negative contrast* effects with rewards. Imagine that an animal is working at a moderate rate for a moderate reward. If the reward is suddenly increased, the animal will work very much harder for a period (perhaps lasting for minutes or longer), but will then gradually revert back to work at a rate close to that at which the animal was working for the moderate reinforcement. This is called positive contrast (see Fig. 2.2). A comparable contrast effect is seen when the reward magnitude (or rate at which rewards are obtained, or the probability of obtaining rewards) is reduced – there is a negative overshoot in the rate of working for a time, but then the rate reverts back to a value close to that at which the animal worked for the moderate reward. This phenomenon is adaptive. It is evidence that animals are in part sensitive to a change in reinforcement, and this helps them to ‘climb reward gradients’ to obtain better rewards. In effect, regardless of the absolute level of reinforcement being received, it is adaptive to be sensitive to a change in reinforcement. If this were not true, an animal receiving very little reinforcement but then obtaining a small increase in positive reinforcement might still be working very little for the reward. But it is much more adaptive to work hard in this situation, as the extra little bit of reward might make the difference between survival or not, and might lead the animal in the direction of even better rewards if what has just been done leads to an improvement in rewards. A similar phenomenon may be evident in humans. People who have very little in the way of rewards, who may be poor, have a poor diet, and may suffer from disease, may nevertheless not have a baseline level of happiness that is necessarily very different from that of a person in an affluent society who in absolute terms apparently has many more rewards. This may be due in part to resetting of the baseline of expected rewards to a constant value, so that we are especially sensitive to changes in rewards (or punishers).

Fifth, in a case where the sight of a stimulus associated with pain produces fear, some philosophers categorize fear as an emotion, but not pain. The distinction they make may be that primary (unlearned) reinforcers do not produce emotions, whereas secondary reinforcers (stimuli associated by stimulus–reinforcement learning with primary reinforcers) do. They describe the pain as a sensation. But neutral stimuli (such as a table) can produce sensations when touched. It accordingly seems to be much more useful to categorize stimuli according to whether they are reinforcing (in which case they produce emotions), or are not reinforcing (in which case they do not produce emotions). Clearly there is a difference between primary reinforcers and learned reinforcers; but this is most precisely caught by noting that this is the difference, and that it is whether a stimulus is reinforcing that determines whether it is related to emotion. Primary and secondary reinforcers have in common that they produce affective states, whereas neutral, non-reinforcing, stimuli do not produce affective states. The major division thus seems to be between stimuli that produce affective states and those that do not; and it is reinforcing stimuli that produce affective states.

Sixth, as we are about to see, emotional states (i.e. those elicited by instrumental reinforcers) have many functions, and the implementations of only some of these functions by the brain are associated with emotional feelings (see Chapters 10 and 4, and Rolls (1999a)). Indeed there is evidence for interesting dissociations in some patients with brain damage between actions performed to reinforcing stimuli and what is subjectively reported. In this sense it is biologically and psychologically useful to consider emotional states to include more than those states associated with feelings of emotion.

Seventh, the role of learning in many emotions should be emphasized. The approach described above shows that the learning of stimulus–reinforcer (i.e. stimulus–reward and stimulus–punisher) associations is the learning involved when emotional responses are learned. In so far as the majority of stimuli that produce our emotional responses do so as a result of learning, this type of learning, and the brain mechanisms that underlie it, are crucial to the majority of our emotions. This, then, provides a theoretical basis for understanding the functions of some brain systems such as the amygdala in emotion, as described in Chapter 4.

It also follows from this approach towards a theory of emotion that brain systems involved in disconnecting stimulus–reinforcer associations when they are no longer appropriate will also be very important in emotion. Failure of this function would be expected to lead, for example, in frustrating situations to inappropriate perseveration of behaviour to stimuli no longer associated with rewards. The inability to correct behaviour when reinforcement contingencies change would be evident in a number of emotion-provoking situations, such as frustration (i.e. non-reward), and the punishment of previously rewarded behaviour. It will be shown in Chapter 4 that this approach, which emphasizes the necessity, in for example social situations, to update and correct the decoded reinforcement value of stimuli continually and rapidly, helps to provide a basis for understanding the functions of some other brain regions such as the orbitofrontal cortex in emotion.

Eighth, understanding the functions of emotion is also important for understanding the nature of emotions, and for understanding the brain systems involved in the different types of response that are produced by emotional states. Emotion appears to have many functions, which are not necessarily mutually exclusive. Some of these functions are described in Chapter

3.

However, a fundamentally important function of emotion that I will propose in Chapter 3 draws out a close link with the definition given here of emotions as states elicited by instrumental reinforcing stimuli, which are the goals for action. I show in Chapter 3 that genes define the goals for (instrumental) actions, and that this is an important Darwinian, adaptive, aspect of brain design. These goals for action are instrumental reinforcers, and this thus helps us to see that by understanding emotions as states elicited by reinforcers, we gain important insight into the nature of emotions. The treatment of the nature of emotion given in this Chapter is thus seen to be directly relevant to understanding this fundamentally important role of emotion in brain design which is related to the role that reinforcers have in guiding actions.

2.5 Summary of the classification of emotion

The theory of emotion outlined above provides systematic and principled ways to categorise different emotions.

One useful way to categorise emotions is to note that the main dimensions of the space of possible emotional responses can well be specified by the different primary reinforcers, examples of which are given in Table 2.1. Within each of these gene-specified dimensions, different reinforcement contingencies would lead to different emotional states, for example pleasure produced by a given taste, and disappointment (frustrative non-reward) if that taste is not available. Also within each of these reinforcer-defined dimensions the exact nature of the primary reinforcing stimulus, including its intensity and variations in its quality (for example in the nature of its texture if it is a somatosensory stimulus), would lead to differences in the emotion elicited. Also within each gene-specified dimension, many states could be cognitively different depending on which different stimulus (e.g. person) had become associated by learning with the primary reinforcer to become a secondary reinforcer. If we then remember that each secondary reinforcer may have many different reinforcement associations, then we see that very large numbers of possible emotions can be described and categorized in this way. Although the description leads to an enormously large numbers of different categories, nevertheless it is systematic, principled, and fairly complete.

Another possible way to categorize emotions might be by reinforcement contingency. For example we might group together into one category all emotions elicited by frustrative non-reward (S_+ and $S_+!$ in Fig. 2.1). Gray (1987) went even further than this, grouping into one category not only the emotions elicited by frustrative non-reward (the non-delivery of rewards), but also those elicited by the delivery of punishers (S_- in Fig. 2.1). Part of his reason for combining these two was that both can lead to decreases in behaviour, which led him to believe that there was a “behavioural inhibition system” (which he identified with the hippocampus) common to both. Clearly at some level the processing is inherently different, in that frustrative non-reward implies a neural system that predicts reward and produces an output if that outcome is not realised, whereas punishment may often involve activation of different sensory systems involved in for example pain. The whole operation and pharmacology of the different circuitry involved in frustrative non-reward and in the effects of punishers must at some level be different, and this is likely to be exploitable in treating emotional states that arise in these two different ways, so it may be very helpful not to combine them into a single category.

In general, categorising emotions by reinforcement contingency alone produces few emotional categories, which is a disadvantage given the many different emotions that can be distinguished, but does have an advantage in producing a rough grouping together of emotional states that do have something in common. However, it should be noted that the reinforcement contingency alone is not a good predictor of the appropriate emotion and emotional behaviour, as shown for example in Fig. 2.1 by the anger that might result from frustrative non-reward if action is possible, and the sadness that might arise if no action is possible to retrieve the lost reward. Further, I note that the axes in Fig. 2.1 refer to only one particular reinforcer (such as a food reward), and are effectively replicated for each different primary reinforcer.

2.6 Other theories of emotion

In the following subsections, I outline some other theories of emotion, and compare them with the above (Rolls') theory of emotion. Surveys of some of the approaches to emotion that have been taken in the past are provided by Strongman (2003) and Oatley and Jenkins (1996).

2.6.1 The James–Lange and other bodily theories of emotion including Damasio's theory

James (1884) believed that emotional experiences were produced by sensing bodily changes, such as changes in heart rate or in skeletal muscles. Lange (1885) had a similar view, although he emphasized the role of autonomic feedback (for example from the heart) in producing the experience of emotion. The theory, which became known as the James–Lange theory, suggested that there are three steps in producing emotional feelings (see Fig. 2.3). The first step is elicitation by the emotion-provoking stimulus of peripheral changes, such as skeleto-muscular activity to produce running away, and autonomic changes, such as alteration of heart rate. But, as pointed out above, the theory leaves unanswered perhaps the most important issue in any theory of emotion: Why do some events make us run away (and then feel emotional), whereas others do not? This is a major weakness of this type of theory. The second step is the sensing of the peripheral responses (e.g. running away, and altered heart rate). The third step is elicitation of the emotional feeling in response to the sensed feedback from the periphery.

The history of research into peripheral theories of emotion starts with the fatal flaw that step one (the question of which stimuli elicit emotion-related responses in the first place) leaves unanswered this most important question. The history continues with the accumulation of empirical evidence that has gradually weakened more and more the hypothesis that peripheral responses made during emotional behaviour have anything to do with producing the emotional behaviour (which has largely already been produced anyway according to the James–Lange theory), or the emotional feeling. Some of the landmarks in this history are as follows.

First, the peripheral changes produced during emotion are not sufficiently distinct to be able to carry the information that would enable one to have subtly different emotional feelings to the vast range of different stimuli that can produce different emotions. The evidence suggests that by measuring many peripheral changes in emotion, such as heart rate, skin conductance, breathing rate, and hormones such as adrenaline and noradrenaline (known in the United States by their Greek names epinephrine and norepinephrine), it may be possible to make coarse distinctions between, for example, anger and fear, but not much finer distinctions (Wagner 1989, Cacioppo, Klein, Berntson and Hatfield 1993, Oatley and Jenkins 1996).

James-Lange theory of emotion

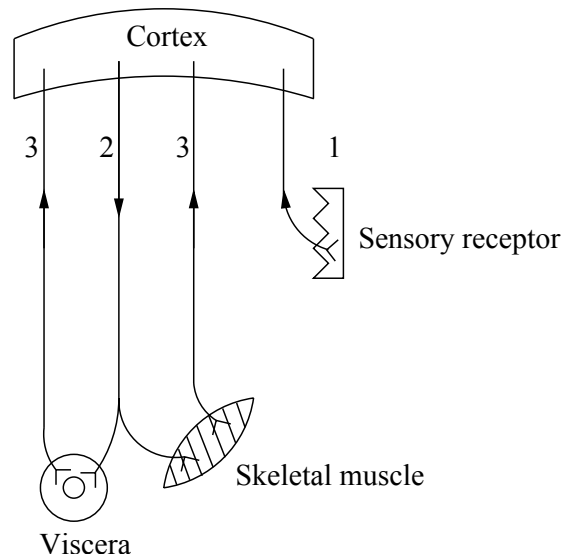


Fig. 2.3 The James–Lange theory of emotion proposes that there are three steps in producing emotional feelings. The first step is elicitation by the emotion-provoking stimulus (received by the cortex via pathway 1 in the Figure) of peripheral changes, such as skeleto-muscular activity to run away, and autonomic changes, such as alteration of heart rate (via pathways labelled 2 in the Figure). The second step is the sensing of the peripheral responses (e.g. altered heart rate, and somatosensory effects produced by running away) (via pathways labelled 3 in the Figure). The third step is elicitation of the emotional feeling in response to the sensed feedback from the periphery.

Second, when emotions are evoked by imagery, then the peripheral responses are much less marked and distinctive than during emotions produced by external stimuli (Ekman, Levenson and Friesen 1983, Stemmler 1989, Levenson, Ekman and Friesen 1990). This makes sense in that although an emotion evoked by imagery may be strong, there is no need to produce strong peripheral responses, because no behavioural responses are required.

Third, disruption of peripheral responses and feedback from them either surgically (for example in dogs, (Cannon 1927, Cannon 1929, Cannon 1931), or as a result of spinal cord injury in humans (Hohmann 1966, Bermond, Fasotti, Niewenhuyse and Schuerman 1991)), does not abolish emotional responses. What was found was that in some patients there was apparently some reduction in emotions in some situations (Hohmann 1966), but this could be related to the fact that some of the patients were severely disabled (which could have produced its own consequences for emotionality), and that in many cases the patients were considerably older than before the spinal cord damage, and this could have been a factor. What was common to both studies was that emotions could be felt by all the patients; and that in some cases, emotions resulting from mental events were even reported as being stronger (Hohmann 1966, Bermond, Fasotti, Niewenhuyse and Schuerman 1991).

Fourth, when autonomic changes are elicited by injections of, for example, adrenaline or noradrenaline, particular emotions are not produced. Instead, the emotion that is produced depends on the cognitive decoding of the reinforcers present in the situation, for example an

actor who insults your parents to make you angry, or an actor who plays a game of hula hoop to make you feel happy (Schachter and Singer 1962). In this situation, the hormone adrenaline or noradrenaline can alter the magnitude of the emotion, but not which emotion is felt. This is further evidence that it is the decoded reinforcement value of the input stimulus or events that determines which emotion is felt. The fact that the hormone injections produced some change in the magnitude of an emotion is not very surprising. If you felt your heart pounding for no explicable reason, you might wonder what was happening, and therefore react more or abnormally.

Fifth, if the peripheral changes associated with emotion are blocked with drugs, then this does not block the perception of emotion (Reisenzein 1983).

Sixth, it is found that in normal life, behavioural expressions of emotion (for example smiling when at a bowling alley) do not usually occur when one might be expected to feel happy because of a success, but instead occur when one is looking at one's friends (Kraut and Johnson 1979). These body responses, which can be very brief, thus often serve the needs of communication, or of action, not of producing emotional feelings.

Despite this rather overwhelming evidence against an important role for body responses in producing emotions or emotional feelings, Damasio (1994) has effectively tried to resurrect a weakened version of the James–Lange theory of emotion from the 19th century, by arguing with his somatic marker hypothesis that after reinforcers have been evaluated, a bodily response ('somatic marker') normally occurs, then this leads to a bodily feeling, which in turn is appreciated by the organism to then make a contribution to the decision-making process⁶. The James–Lange theory has a number of major weaknesses just outlined that apply also to the somatic marker hypothesis.

The somatic marker hypothesis postulates that emotional decision-making is facilitated by peripheral feedback from for example the autonomic nervous system. In a direct test of this, Heims, Critchley, Dolan, Mathias and Cipolotti (2004) measured emotional decision making using the Iowa Gambling Task (Bechara, Damasio, Damasio and Anderson 1994, Bechara, Tranel, Damasio and Damasio 1996, Bechara, Damasio, Tranel and Damasio 1997, Damasio 1994) (described in Section 4.5.6) in patients with pure autonomic failure. In this condition, there is degeneration of the peripheral autonomic system, and thus autonomic responses are severely impaired, and there can be no resulting feedback to the brain. It was found that performance in the Iowa Gambling Task was not impaired, and nor were many other tests of emotion and emotional performance, including face expression identification, theory of mind tasks of social situations, and social cognition tasks. Thus emotional decision-making does not depend on the ongoing feedback from somatic markers related to autonomic function. Damasio might argue that feedback from the autonomic system is not actually important, and that it is feedback from skeletomotor responses such as arm movements or muscle tension that is important. He might also argue that the autonomic feedback is not usually necessary for emotional decision making, because it can be 'simulated' by the rest of the brain. However,

⁶In the James–Lange theory, it was emotional feelings that depend on peripheral feedback; for Damasio, it is the decision of which behavioural response to make that is normally influenced by the peripheral feedback. A quotation from Damasio (1994, p190) follows: "The squirrel did not really think about his various options and calculate the costs and benefits of each. He saw the cat, was jolted by the body state, and ran." Here it is clear that the pathway to action uses the body state as part of the route. Damasio would also like decisions to be implemented using the peripheral changes elicited by emotional stimuli. Given all the different reinforcers that may influence behaviour, Damasio (1994) even suggests that the net result of them all is reflected in the net peripheral outcome, and then the brain can sense this net peripheral result, and thus know what decision to take.

the study by Heims et al. (2004) does show that ongoing autonomic feedback is not necessary for normal emotional decision-making, and this leaves the somatic marker hypothesis more precarious.

Part of the evidence for the somatic marker hypothesis was that normal participants in the Iowa Gambling Task were described as deciding advantageously before knowing the advantageous strategy (Bechara, Damasio, Tranel and Damasio 1997). The interpretation was that they had implicit (unconscious) knowledge implemented via a somatic marker process that was used in the task, which was not being solved by explicit (conscious) knowledge. Maia and McClelland (2004) (see also Maia and McClelland (2005)) however showed that with more sensitive questioning, normal participants at least had available to them explicit knowledge about the outcomes of the different decks that was as good as or better than the choices made, weakening the arguments of Bechara et al. (1997) and Bechara, Damasio, Tranel and Damasio (2005) that the task was being solved implicitly and using somatic markers. Further evidence on factors that contribute to the effects found in the Iowa Gambling Task are described in Section 4.5.6.

Another argument against the somatic marker hypothesis is that there can be dissociations between autonomic and other indices of emotion, thus providing evidence that behaviour may not follow from autonomic and other effects. For example, lesions of different parts of the amygdala influence autonomic responses and instrumental behaviour differently, as shown in Section 4.6.3 and Fig. 4.52.

Another major weakness, which applies to both the James–Lange and to Damasio’s somatic marker hypothesis, is that they do not take account of the fact that once an information processor has determined that a response should be made or inhibited based on reinforcement association, a function attributed in the theory proposed in this Chapter and by Rolls (Rolls 1986c, Rolls 1986a, Rolls 1990d, Rolls 1999a) to the orbitofrontal cortex, it would be very inefficient and noisy to place in the execution route a peripheral response, and transducers to attempt to measure that peripheral response, itself a notoriously difficult procedure (see, e.g., Grossman (1967)). Even for the cases when Damasio (1994) might argue that the peripheral somatic marker and its feedback can be by-passed using conditioning of a representation in, e.g., the somatosensory cortex to a command signal (which might originate in the orbitofrontal cortex), he apparently would still wish to argue that the activity in the somatosensory cortex is important for the emotion to be appreciated or to influence behaviour. (Without this, the somatic marker hypothesis would vanish.) The prediction would apparently be that if an emotional response were produced to a visual stimulus, then this would necessarily involve activity in the somatosensory cortex or other brain region in which the ‘somatic marker’ would be represented. This prediction could be tested (for example in patients with somatosensory cortex damage), but it seems most unlikely that an emotion produced by a visual reinforcer would require activity in the somatosensory cortex to feel emotional or to elicit emotional decisions. However, Adolphs, Tranel and Denburg (2000) have pursued this general line of enquiry, and report that the more damage there is to somatosensory cortex, the greater the impairment in the emotional state reported by patients. However, the parts of the somatosensory system that appear to be damaged most frequently in the patients with emotional change are often in the anterior and ventral extensions of the somatosensory cortex in insular and nearby areas, and it would be useful to know whether this damage interrupted some of the connections or functions of the orbitofrontal cortex areas just anterior.

More recently, Damasio has stated the somatic marker hypothesis in a weak form, suggesting that somatic markers do not even reflect the valence of the reinforcer, but just provide a signal that depends on the intensity of the emotion, independently of the type of emotion. On this view, the role of somatic markers in decision making would be very general, providing, as Damasio says, just a jolt to spur the system on (A.R.Damasio, paper delivered at the 6th Annual Wisconsin Symposium on Emotion, April 2000).

The alternative view proposed here (and by Rolls (1986c), Rolls (1986a), Rolls (1990d), Rolls (1999a), and Rolls (2000f)) is that where the reinforcement value of the visual stimulus is decoded, namely in the orbitofrontal cortex and the amygdala, is the appropriate part of the brain for outputs to influence behaviour (via, e.g., the orbitofrontal-to-striatal connections), and that the orbitofrontal cortex and amygdala, and brain structures that receive connections from them, are the likely places where neuronal activity is directly related to emotional states and to felt emotions (see further Chapter 10 and Rolls (1999a)).

2.6.2 Appraisal theory

Appraisal theory, developed and described by Frijda (1986), Oatley and Johnson-Laird (1987), Lazarus (1991), Izard (1993), Stein, Trabasso and Liwag (1994), Oatley and Jenkins (1996), and Scherer (1999) (see also Scherer (2001) and Scherer, Schorr and Johnstone (2001)) generally holds that two types of appraisal are involved in emotion. Primary appraisal holds that “an emotion is usually caused by a person consciously or unconsciously evaluating an event as relevant to a concern (a goal) that is important; the emotion is felt as positive when a concern is advanced and negative when a concern is impeded” (from Oatley and Jenkins (1996), p. 96). As noted above, the concept of appraisal presumably involves assessment of whether something is a reward or punisher, that is whether it will be worked for or avoided. The description in terms of rewards and punishers adopted here simply seems much more precisely and operationally specified. If primary appraisal is defined with respect to goals, it might be helpful to note that goals may just be the reinforcers specified in Rolls’ theory, and if so the reinforcer/punisher approach provides clear definitions of goals (as reinforcers, see Appendix 3), which is helpful, precise, and makes a link to what may be specified by genes.

Secondary appraisal is concerned with coping potential, that is with whether for example a plan can be constructed, and how successful it is likely to be.

I note that appraisal theory is in many ways quite close to the theory that I outline here and in *The Brain and Emotion* (Rolls 1999a), and I do not see them as rivals. Instead, I hope that those who have an appraisal theory of emotion will consider whether much of what is encompassed by primary appraisal is not actually rather close to assessing whether stimuli or events are reinforcers; and whether much of what is encompassed by secondary appraisal is rather close to taking into account the actions that are possible in particular circumstances, as described above in Section 2.2.

An aspect of some flavours of appraisal theory with which I do not agree is that emotions have as one of their functions releasing particular actions, which seems to make a link with species-specific action tendencies or responses (Tomkins 1995, Panksepp 1998) or more ‘open motor programs’ (Ekman 2003). I argue in Chapter 3 that rarely are behavioural responses programmed by genes (see Table 2.1), but instead genes optimise their effects on behaviour if they specify the goals for (flexible) actions, that is if they specify rewards and punishers. The difference is quite considerable, in that specifying goals is much more economical in terms of the information that must be encoded in the genome; and in that specifying goals for actions

allows much more flexibility to the actual actions that are produced. Of course I acknowledge that there is some preparedness to learn associations between particular types of secondary and primary reinforcers, and see this just as an economy of sensory–sensory convergence in the brain, whereby for example it does not convey much advantage to be able to learn that flashing lights (as contrasted with the taste of a food just eaten) are followed by sickness.

2.6.3 Dimensional and categorical theories of emotion

These theories suggest that there are a number of fundamental or basic emotions. Charles Darwin for example in his book *The Expression of the Emotions in Man and Animals* (1872) showed that some basic expressions of emotion are similar in animals and humans. Some of the examples he gave are shown in Table 2.1. His focus was on the continuity between animals and humans of how emotion is expressed.

In a development of this approach, Ekman and colleagues (Ekman 1982, Ekman 1992, Ekman 1993, Ekman, Friesen and Ellsworth 1972, Ekman, Levenson and Friesen 1983) have suggested that humans categorize face expressions into a number of basic categories that are similar across cultures. These face expression categories include happy, fear, anger, surprise, grief and sadness.

A related approach is to identify a few clusters of variables or factors that result from multidimensional analysis of questionnaires, and to identify these factors as basic emotions. (Multidimensional analyses such as factor analysis seek to identify a few underlying sources of variance to which a large number of data values such as answers to questions are related.) The categories of emotions identified in these ways may be supported by correlating them with autonomic measures (e.g. Ekman et al. (1983)).

One potential problem with some of these approaches is that they risk finding seven plus or minus two categories, which is the normal maximal number of categories with which humans normally operate, as described in a famous paper by George Miller (1956). A second problem is that there is no special reason why the first few factors (which account for most of the variance) in a factor analysis should provide a complete or principled classification of different emotions, or of their functions. In contrast, the theory described here does produce a principled classification of different emotions based on reinforcement contingencies, the nature of the primary and secondary reinforcers, etc, as set out in Sections 2.2 and 2.3. Moreover, the present theory links the functions of emotions to the classification produced, by showing how the functions of emotion can be understood in terms of the gene-specified reinforcers that produce different emotions (see Chapter 3).

An opposite approach to the dimensional or categorical approach is to attempt to describe the richness of every emotion (e.g. Ben-Ze'ev (2000)). Although it is important to understand the richness of every emotion, I believe that this is better performed with a set of underlying principles of the type set out above (in Section 2.2), rather than without any obvious principles to approach the subtlety of emotions.

2.6.4 Other approaches to emotion

LeDoux (LeDoux 1992, LeDoux 1995, LeDoux 1996) has described a theory of the neural basis of emotion that is probably conceptually similar to that of Rolls (Rolls 1975, Rolls 1986c, Rolls 1986a, Rolls 1990d, Rolls 1995b, Rolls 1999a, Rolls 2000f) (and this book), except that he focuses mostly on the role of the amygdala in emotion (and not on other brain

regions such as the orbitofrontal cortex, which are poorly developed in the rat); except that he focuses mainly on fear (based on his studies of the role of the amygdala and related structures in fear conditioning in the rat); and except that he suggests from his neurophysiological findings that an important route for conditioned emotional stimuli to influence behaviour is via the subcortical inputs (especially auditory from the medial part of the medial geniculate nucleus of the thalamus) to the amygdala. This theory is discussed further on pages 169–170.

Panksepp's (1998) approach to emotion has its origins in neuroethological investigations of brainstem systems that when activated lead to behaviours like fixed action patterns, including escape, flight and fear behaviour. His views about consciousness include the postulate that "feelings may emerge when endogenous sensory and emotional systems within the brain that receive direct inputs from the outside world as well as the neurodynamics of the SELF (a Simple Ego-type Life Form) begin to reverberate with each other's changing neuronal firing rhythms" (Panksepp 1998) (p. 309).

Other approaches to emotion are summarized by Strongman (2003).

2.7 Individual differences in emotion, personality, and emotional intelligence

Hans J. Eysenck developed the theory that personality might be related to different aspects of conditioning. He analysed the factors that accounted for the variance in the differences between the personality of different humans (using, for example, questionnaires), and suggested that the first two factors in personality (those which accounted for most of the variance) were introversion vs extraversion, and neuroticism (related to a tendency to be anxious). He performed studies of classical conditioning on groups of subjects, and also obtained measures of what he termed arousal. Based on the correlations of these measures with the dimensions identified in the factor analysis, he suggested that introverts showed greater conditionability (to weak stimuli) and are more readily aroused by external stimulation than extraverts; and that neuroticism raises the general intensity of emotional reactions (see Eysenck and Eysenck (1968) and Eysenck and Eysenck (1985)).

Jeffrey A. Gray (1970) reinterpreted the findings, suggesting that introverts are more sensitive to punishment and frustrative non-reward than are extraverts; and that neuroticism reflects the extent of sensitivity to both reward and punishment (see Matthews and Gilliland (1999)). A related hypothesis is that extraverts may show enhanced learning in reward conditions, and may show enhanced processing of positively valent stimuli (Rusting and Larsen 1998). Matthews and Gilliland (1999), reviewing the evidence, show that there is some support for both hypotheses about introversion vs extraversion, namely that introverts may in general condition readily, and that extraverts may be relatively more responsive to reward stimuli (and correspondingly, introverts to punishers). However, Matthews and Gilliland (1999) go on to show that extraverts may perform less well at vigilance tasks (in which the subject must detect stimuli that occur with low probability); may tend to be more impulsive; and perform better when arousal is high (e.g. later in the day), and when rapid responses rather than reflective thought is needed (see also Matthews, Zeidner and Roberts (2002)). With respect to neuroticism and trait anxiety, anxious individuals tend to focus attention on potentially threatening information (punishers) at the cost of neglecting neutral or positive information sources; and may make more negative judgements, especially in evaluating self-worth and personal competence (Matthews, Zeidner and Roberts 2002).

More recent evidence comes from recent functional neuroimaging studies. For example, Canli, Sivers, Whitfield, Gotlib and Gabrieli (2002) have found that happy face expressions are more likely to activate the human amygdala in extraverts than in introverts. In addition, positively affective pictures interact with extraversion, and negatively affective pictures with neuroticism to produce activation of the amygdala (Canli, Zhao, Desmond, Kang, Gross and Gabrieli 2001, Hamann and Canli 2004). This supports the conceptually important point made above that part of the basis of personality may be differential sensitivity to different rewards and punishers, and omission and termination of different rewards and punishers.

The observations just described are consistent with the hypothesis that part of the basis of extraversion is increased reactivity to positively affective (as compared to negatively affective) face expressions and other positively affective stimuli including pictures. The exact mechanisms involved may be revealed in the future by genetic studies, and these might potentially address for example whether genes control responses to positively affective stimuli, or whether some more general personality trait by altering perhaps mood produces differential top-down biasing of face expression decoding systems in the way outlined in Section 4.10.

Another example is the impulsive behaviour that is a part of Borderline Personality Disorder (BPD), which could reflect factors such as less sensitivity to the punishers associated with waiting for rational processing to lead to a satisfactory solution, or changes in internal timing processes that lead to a faster perception of time (Berlin, Rolls and Kischka 2004, Berlin and Rolls 2004) (see Section 4.5.6). It was of considerable interest that the BPD group (mainly self-harming patients), as well as a group of patients with damage to the orbitofrontal cortex, scored highly on a Frontal Behaviour Questionnaire that assessed inappropriate behaviours typical of orbitofrontal cortex patients including disinhibition, social inappropriateness, perseveration, and uncooperativeness. In terms of measures of personality, using the Big Five personality measure, both groups were also less open to experience (i.e. less open-minded). In terms of other personality measures and characteristics, the orbitofrontal and BPD patients performed differently: BPD patients were less extraverted and conscientious and more neurotic and emotional than the orbitofrontal group (Berlin, Rolls and Kischka 2004, Berlin and Rolls 2004, Berlin, Rolls and Iversen 2005). Thus some aspects of personality, such as impulsiveness and being less open to experience, but not other aspects, such as extraversion, neuroticism and conscientiousness, were differentially related to orbitofrontal cortex function.

Daniel Goleman (1995) has popularized the concept of *emotional intelligence*. The rather sweeping definition given was “Emotional intelligence [includes] abilities such as being able to motivate oneself and persist in the face of frustrations, to control impulse and delay gratification; to regulate one’s moods and keep distress from swamping the ability to think; to empathize and to hope” (Goleman (1995), p. 34).

One potential problem with this definition of emotional intelligence as an ability is that different aspects within this definition (such as impulse control and hope) may be unrelated, so a unitary ability described in this way seems unlikely. An excellent critical evaluation of the concept has been produced by Matthews, Zeidner and Roberts (2002). They note (p. 368) that in a rough and ready way, one might identify personality traits of emotional stability (low neuroticism), extraversion, agreeableness, and conscientiousness/self-control as dispositions that tend to facilitate everyday social interaction and to promote more positive emotion. (Indeed, one measure of emotional intelligence, the EQ-i (Bar-On 1997), has high correlations with some of the Big Five personality traits, especially, negatively, with neuroticism, and the EQ-i may reflect three constructs, self-esteem, empathy, and impulse control (Matthews et

al. 2002).) But these personality traits are supposed to be independent, so linking them to a single ability of emotional intelligence is inconsistent. Moreover, this combination of personality traits might well not be adaptive in many circumstances, so the concept of this combination as an ‘ability’ is inappropriate (pp. 368–370).

However, the concept of emotional intelligence does appear to be related in a general way to the usage of the (mainly clinical) term ‘alexithymia’, in a sense the opposite, which includes the following components: (a) difficulty in identifying and describing emotions and distinguishing between feelings and the bodily sensations of arousal, (b) difficulty in describing feelings to other people, (c) constricted imaginal processes, as evidenced by a paucity of fantasies, and (d) a stimulus-bound externally oriented cognitive style, as evidenced by preoccupation with the details of external events rather than inner emotional experiences (Matthews et al. 2002). In terms of personality, alexithymia converges with the first three dimensions of the Five Factor Model of personality (FFM, the Big Five model), with high N (vulnerability to emotional distress), low E (low positive emotionality), and a limited range of imagination (low O) (Matthews et al. 2002). Indeed, alexithymia is strongly inversely correlated with measures of emotional intelligence, suggesting that emotional intelligence may be a new term that encompasses much of the opposite of what has been the important concept of alexithymia in the clinical literature for more than 20 years (Matthews et al. 2002). Alexithymics have difficulties in identifying face expressions (Lane, Sechrest, Reidel, Weldon, Kaszniak and Schwartz 1996), suggesting some impairments in the fundamental processing of emotion-related information, in particular capacities known to require the orbitofrontal and anterior cingulate cortices (Hornak, Bramham, Rolls, Morris, O’Doherty, Bullock and Polkey 2003). Consistently, it has been found that anterior cingulate cortex activation is correlated across individuals with their ability to recognize and describe emotions induced either by films or by the recall of personal experiences (Lane, Reiman, Axelrod, Yun, Holmes and Schwartz 1998). In summary, emotional intelligence, and what is largely its opposite, alexithymia, is probably not a particular ability, is not independent of existing personality measures, but does encompass a number of probably different ways in which individuals may differ in their emotion-related processing (Rolls 2007c).

I do not consider this research area in much more detail. However, I do point out that insofar as sensitivity to rewards and punishers, and the ability to learn and be influenced by rewards and punishers, may be important in personality, and are closely involved in emotion according to the theory developed here, there may be close links between the neural bases of emotion, to be described in Chapter 4, and personality. An extreme example might be that if humans were insensitive to social punishers following orbitofrontal cortex damage, we might expect social problems, and indeed Tranel, Bechara and Denburg (2002) have used the term ‘acquired sociopathy’ to describe some of these patients.

More generally, we might expect sensitivity to different types of reinforcer (including social reinforcers) to vary between individuals both as a result of gene variation and as a result of learning, and this, operating over a large number of different social reinforcers, might produce many different variations of personality based on the sensitivity to a large number of different reinforcers. Further, insofar as the functions of particular brain regions may be related to particular processes involved in emotion [with evidence for example that the human orbitofrontal cortex is involved in face expression decoding, and in impulsiveness, but not in some other aspects of personality (see Section 4.5.6)], then it may be possible in future to understand different particular modules for inter-relations between reward/punishment and

personality systems.

The concept of the relation between differential sensitivity to different types of reward and punisher might produce individuals showing many types of conditional evolutionarily stable strategies (see footnote 19 on page 358), where the conditionality of the strategy might be influenced in different individuals by differential sensitivity to different rewards and punishers. Examples of behaviours that might be produced in this way are included in Chapter 9.

2.8 Cognition and Emotion

It may be noted that while the definition of emotions as states elicited by reinforcers (with particular functions) is operational, it should not be criticized as behaviourist (Katz 2000). For example, the definition has nothing to do with stimulus–response (habit) associations, but instead with a two-stage type of learning, in which a first stage is learning which environmental stimuli or events are associated with reinforcers, which potentially is a very rapid and flexible process; and a second stage produces appropriate instrumental and arbitrary actions performed in order to achieve the goal (which might be to obtain a reward or avoid a punisher). In the instrumental stage, animals learn about the outcomes of their actions (see Dickinson (1994), Pearce (1997)).

To determine what is a goal for an action, every type of cognitive operation may be involved. The proposal is that whatever cognitive operations are involved, then if the outcome is that a certain event, stimulus, thought (or any one of these remembered) leads to the evaluation that the event is a reward or punisher, then an emotion will be produced. So cognition is far from excluded.

Indeed, cognitive operations may produce emotions when operating at three levels of the architecture, as described more fully in Chapter 3. The first is the implicit level (see Fig. 10.4), where a primary reinforcer, or a stimulus or event associated with a primary reinforcer, may lead to emotions. The second level is where a (first order) syntactic symbol processing system performing “what ... if” computations to implement planning results in identification of a rewarding or punishing outcome. The third level is the higher order linguistic thought level described in Chapter 10, where thinking about and evaluating the operations of a first order linguistic processor may result in a reinforcing outcome such as “I should not spend further time thinking about that set of plans, as it would be better now to devote my linguistic resources (which are limited and serial) to this other set of plans”.

Another way in which cognition influences emotion is that cognitive states, even at the level of language, can modulate subjective and brain responses to affective stimuli, as analysed in Section 4.5.5.7. There an experiment is described in which a word label (‘cheese’ vs ‘body odour’) influences the pleasantness ratings, and the activations in olfactory stages at least as early as the secondary olfactory cortex in the orbitofrontal cortex, to a standard test odour (De Araujo, Rolls, Velazco, Margot and Cayeux 2005). An implication of these findings is that language-based cognitive states can influence even relatively early cortical representations of rewards and punishers, and thus potentially modulate how much emotion is felt subjectively to an emotion-provoking stimulus.

I suggest that this top-down modulation occurs in a way that is exactly analogous to top-down attentional effects, which are believed to be implemented by a top-down biased competition mechanism (Rolls and Deco 2002, Deco and Rolls 2003, Deco and Rolls 2005b, Rolls and Stringer 2001b). In this case, the semantic, language-based, representation is the

source of the biased competition, and the effect could be not only to bias the early cortical representation of a reward or punisher in one direction or another, but also by providing much or little top-down modulation, to influence how much emotion is felt (see Chapter 10), by modulating the processing of emotion-related stimuli (including remembered stimuli or events) at relatively early processing stages. This could be a mechanism by which cognition can influence how much emotion is felt under conditions in which emotions such as empathy and pity may occur, and when for example reading a novel, attending a play, listening to music, etc (see Section 11.4). Analysis of the mechanisms by which the top-down biased competition operates are becoming detailed (Desimone and Duncan 1995, Rolls and Deco 2002, Deco and Rolls 2003, Deco and Rolls 2004, Deco and Rolls 2005b), and are included in the model described in Appendix 2 in which a rule module exerts a top-down influence on neurons that represent stimulus–reward and stimulus–punisher combinations to influence which stimulus should currently be interpreted as reward-related (see also Deco and Rolls (2005d)).

Another way in which cognitive factors are related to emotion is that mood can affect cognitive processing, and one of the effects of this is to promote continuity of behaviour (see Chapter 3). One of the mechanisms described (in Section 4.10) utilizes backprojections to cortical areas from the amygdala and orbitofrontal cortex, so that reciprocal interactions between cognition and emotion are made possible.

2.9 Emotion, motivation, reward, and mood

It is useful to be clear about the difference between motivation, emotion, reward, and mood (cf. Rolls (2000f)). **Motivation** makes one work to obtain a reward, or work to escape from or avoid a punisher. One example of motivation is hunger, and another thirst, which in these cases are states set largely by internal homeostatically-related variables such as plasma glucose concentration and plasma osmolality. A reward is a stimulus or event that one works to obtain, such as food, and a punisher is what one works to escape from or avoid (or which suppresses an action on which its delivery is contingent), such as a painful stimulus or the sight of an object associated with a painful stimulus. Obtaining the reward or avoiding the punisher is the goal for the action. A motivational state is one in which a goal is *desired*. An **emotion** is a state elicited when a goal is obtained, that is by an instrumental reinforcer (i.e. a reward or punisher, or omission or termination of a reward or punisher), for example fear produced by the sight of the object associated with pain. This makes it clear that emotions are states elicited by rewards or punishers that have particular functions.

Of course, one of the functions of emotions is that they are motivating, as exemplified by the case of the fear produced by the sight of the object that can produce pain, which motivates one to avoid receiving the painful stimulus, which is the goal for the action. In that emotion-provoking stimuli or events produce motivation, then arousal is likely to occur, especially for reinforcers that lead to the active initiation of actions. However, arousal alone is not sufficient to define motivation or emotion, in that the motivational state must specify the particular type of goal that is the object of the motivational state, such as water if we are thirsty, food if we are hungry, and avoidance of the painful unconditioned stimulus signalled by a fear-inducing conditioned stimulus.

A **mood** is a continuing state normally elicited by a reinforcer, and is thus part of what is an emotion. The other part of an emotion is the decoding of the stimulus in terms of whether it is a reward or punisher, that is, of what causes the emotion, or in philosophical

terminology of what the emotion is about or the object of the emotion. Mood states help to implement some of the persistence-related functions of emotion, can continue when the originating stimulus may be forgotten (by the explicit system described in Chapter 10), and may occur spontaneously not because such spontaneous mood swings may have been selected for, but because of the difficulty of maintaining stability of the neuronal firing that implements mood (or affective) state (see *The Brain and Emotion*, Rolls (1999a), pp. 62, 66). Mood states are thus not necessarily about an object.

Thus, motivation may be seen as a state in which one is working for a goal, and emotion as a state that occurs when the goal, a reinforcer, is obtained, and that may persist afterwards. The concept of gene-defined reinforcers providing the goals for action helps to understand the relation between motivational states (or desires) and emotion, as the organism must be built to be motivated to obtain the goals, and to be placed in a different state (emotion) when the goal is or is not achieved by the action. Emotional states may be motivating, as in frustrative non-reward. The close but clear relation between motivation and emotion is that both involve what humans describe as affective states (e.g. feeling hungry, liking the taste of a food, feeling happy because of a social reinforcer), and both are about goals. The Darwinian theory of the functions of emotion developed in Chapter 3 which shows how emotion is adaptive because it reflects the operation of a process by which genes define goals for action applies just as much to motivation (see further Section 3.6), in that emotion can be thought of as states elicited by goals (rewards and punishers), and motivation can be thought of as states elicited when goals are being sought. By specifying goals the genes must specify both that we must be motivated to obtain those goals, and that when the goals are obtained, further states, emotional states with further functions, are produced.

2.10 Is the concept of emotion still useful when we understand its mechanisms?

Kralik and Hauser (2000) ask whether it is helpful to maintain the concept of an emotional state when one starts to understand the mechanisms of reward and punisher decoding, the selection of actions, etc. My view is that emotion is a helpful concept, for a number of reasons.

First, the state is produced by clearly defined stimuli (see above).

Second, the state has many different functions, summarized in Chapter 3, so that a model in which a stimulus is connected to a single output is inappropriate. In these circumstances, an intervening state that implements many functions is useful.

Third, one of the functions of emotion is to support the selection of any appropriate action to a reward or punisher, or its omission or termination, as in two-process learning. In the first stage, an emotional state is produced, and in the second stage, any action is selected that is appropriate given the emotional state. For example, if fear is the emotional state produced by a pain-associated stimulus, an action will be selected to escape from or avoid the emotion-provoking stimulus. In that emotion is a state that guides the elicitation of an action to a stimulus, the emotional state is not itself a behavioural response.

Fourth, other functions of emotional states include the biasing of cognitive function to influence the interpretation of future events, which is clearly not a response.

Fifth, emotional states have the important properties that they persist for times in the order of minutes or hours, thus maintaining persistence of behaviour and consistency of action even after the emotion-provoking stimulus has disappeared.

Sixth, the concept of emotional states just described maps neatly onto folk-psychological concepts of emotions, and provides a convenient conceptual level that bridges to the low-level description of exactly how the stimuli are decoded to elicit the state, how the state is maintained, and how it performs its many functions.

The concept of an emotional state is thus clearly defined in terms of how stimuli elicit the state, and of the many functions of the state including the selection of action. Emotional states are not the stimuli themselves, nor the stimulus decoding, nor the responses finally selected, but consist of on-going states elicited by stimuli in the way described, and performing the functions described. We are indeed starting to understand how the different types of processing involved are implemented in the brain, and these are some of the types of advance described in *The Brain and Emotion* (Rolls 1999a) and in this book. But understanding the implementation of the processes involved in emotion does not mean that emotion itself as a useful concept at its own level will disappear.

In addition, understanding ‘how’ emotion works will not address a number of important questions about emotion, including the ‘why’ questions about for example the evolutionary adaptive value of emotions (see Section 3.1).

2.11 Advantages of the approach to emotion described here (Rolls’ theory of emotion)

I now evaluate the advantages of and justifications for starting with the concept that emotions are states elicited by instrumental reinforcers, even though one proposes that a full definition requires the principles summarized in section 2.2, and incorporating a statement of the functions elicited by those states.

One advantage is that this definition in terms of rewards and punishers may provide a concise operational definition of the environmental stimuli or events that actually lead to emotions. If we can agree that the environmental conditions that lead to emotions are those that can be described as rewarding or punishing, and that those that are not rewards or punishers do not lead to states that are described as emotional, then we are a long way forward in producing a conceptualization of what emotions may be. No commentators on the *Précis of The Brain and Emotion* (Rolls 2000f) actually produced clear exceptions to this correspondence. If we accept this operational definition, it provides us with a powerful way forward to start to examine emotions (because we accept that they are states elicited by rewards or punishers, and have a useful delimitation of what events produce emotion). This leads directly to an analysis of the brain mechanisms that implement emotions as those brain mechanisms that decode environmental stimuli as primary reinforcers, those brain mechanisms that implement stimulus–reinforcer association learning, and the brain mechanisms that link the resulting emotional states to actions.

A second advantage of this definition is that it enables us to see emotions in the context of what I propose is their most important function, namely as a way to provide a mechanism for the genes to influence behaviour in a brain that evolves by gene selection. It is argued in Chapter 3 that the genes do this by specifying the stimuli or events that the animal is built to find rewarding or punishing, i.e. to find reinforcing, so that the genes specify the goals for action, not the actions themselves. The definition of emotions as states elicited by reinforcers thus links directly to the Darwinian theory I propose for why we have emotions, which is that some genes specify reinforcers, that is goals for action, that will increase the fitness of these

genes. It is these particular genes that specify reinforcers that provide the foundation I propose for emotional states. The definition of emotion in terms of states elicited by reinforcers should not be seen thus as behaviourist, but instead as part of a much broader theory that takes an adaptive, Darwinian, approach to the functions of emotion, and how they are important in brain design (see Section 3.5).

A third advantage is that the definition offers a principled way to approach emotion. Different emotions can be classified and understood in terms of different reinforcement contingencies and different reinforcers, and hence directly in terms of their functions. This is recommended as being more advantageous than categorizing emotions based on clusters of variables or factors that result from multidimensional analysis of questionnaires etc, or by correlation with autonomic or face expression measures, which do not lead directly to an understanding of the different functions of different emotions (and run the risk of producing seven plus or minus two categories, cf. Miller (1956)), as described in section 2.6.3. This definition of emotion also leads to an operational, and thus clearly specified, approach to emotions, whereas approaches such as appraisal theory may suffer from the disadvantage that they quickly become somewhat under-specified and intractable, as described in section 2.6.2. Moreover, this principled way of understanding emotions provides a systematic and fundamental way to approach the brain mechanisms involved in emotion, in that brain regions involved in decoding primary reinforcers, and brain regions involved in learning associations of events to primary reinforcers, can be seen to have a clear information-processing role in emotion. Analysing the information processing performed by each connected stage in the brain provides a fruitful approach to understanding neural computation (Rolls and Treves 1998, Rolls and Deco 2002).

In the context of emotion, this approach is also more principled and systematic than identifying categories of (sometimes ethologically described) behaviour such as playfulness and aggression, and looking for brain centres specialized for each category of behaviour (cf. Panksepp (1998)). The specification of actions such as fixed action patterns (in contrast to goals) by genes is not only genetically expensive, but having brain regions specialized for actions (such as playfulness and rage, cf Panksepp (1998)) would lead to a multitude of specialized brain action/emotion systems, with potentially one for every possible type of emotional response. In contrast, specifying emotions as states elicited by reinforcers leaves open and flexible the particular action that may be taken in particular circumstances, and has the great advantage of economy of genetic specification (the genes need only specify what is rewarding and punishing). (Of course, as described above, the type of coping by actions that is possible may influence the emotional state, as in the case of sadness vs anger.)

Specifying emotions in terms of the types of rewards and punishers that elicit the emotion may of course also lead to spatially separated brain systems especially involved in different types of emotion, for the primary reinforcers (such as taste, touch, pain, the failure to receive an expected reward, or a face expression, and learning about these reinforcers) may be decoded and represented in different brain regions. This is because of their different input pathways to the brain, and the utility of forming representations to the object level within each sensory modality before reward value is made explicit in the representation, leading to some specialization of different brain regions and systems in different types of emotion (see Chapter 4).

A fourth advantage of conceptualizing emotions as states elicited by reinforcers is that this provides an immediate way into understanding the relation between emotion and personality

(see section 2.7).

A complex issue related to one's definition of emotion is where the boundaries for emotional states should be set. Should our definition result in emotions being states that occur in invertebrates such as *Aplysia*, as suggested by Kupferman (2000)? My own answer to this is to set off from emotions those behaviours that are performed with fixed responses, that is without the possibility for selecting arbitrary types of behaviour as the goals for actions (see Chapter 3). Such fixed-response behaviours include taxes, such as might be performed by a single cell organism swimming up a chemical gradient towards a source of nutrient. One reason why these types of behaviour with fixed responses are excluded from emotion (though they may be forerunners to it) is that the behaviour does not occur by elicitation of a persistent or continuing state to a reinforcing stimulus that provides the motivation for (arbitrary) instrumental responses to obtain the goal. (That an instrumental (or operant) response is being made is demonstrated most precisely by the bidirectional criterion that either a response, or its opposite, may be performed as an action to obtain a goal.) It is the intervening persistent state elicited by reinforcing stimuli and the ability to allow stimuli to be interfaced to arbitrary instrumental responses or actions that is one of the prime functions of emotion described here (see Chapter 3), and is therefore incorporated into the definition of emotion. The definition thus provides a clear way of dividing states into emotional or not, as it includes only those states that allow instrumental learning, that is arbitrary actions to be performed to obtain reinforcing outcomes (such as obtaining rewards and avoiding punishers). Although animals that do not perform instrumental learning may not qualify according to this criterion as having emotions, they may of course have states that are precursors to emotions. This discussion thus leads to one possible way to separate animals that have emotions from those that do not, a way that is related to one of the fundamental functions of emotion, but it is realised that the separation made at this point should be seen as a useful separating point with a clear principle underlying it, but not a separating point that need be thought of as more than a useful convention in this context.