1 Reinforcement Sensitivity Theory (RST): introduction

Philip J. Corr

The *Reinforcement Sensitivity Theory* (RST) of personality is a theoretical account of the neural and psychological processes underlying the major dimensions of personality. The first of this introductory chapter traces the development of RST, from its official birth in 1970, through to Gray's highly influential 1982 *The Neuropsychology of Anxiety*, and on to its major revision in 2000 with the second edition of this book (co-authored with Neil McNaughton) – this Part may be read as an overview tutorial of RST. The second discusses some of the major issues facing future RST research. The third turns attention to the question of the level of behavioural control exerted by 'biological' and 'cognitive' processes, and discusses the implications of findings from consciousness studies for conceptualizing the role of these processes in RST.

Past and present

At the time of writing (2006), most empirical studies continue to test the unrevised (pre-2000) version of RST. But, in many crucial respects, the revised (2000) theory is very different, leading to the formulation of new hypotheses, some of which stand in opposition to those generated from the unrevised theory. This reluctance, or slowness, to adopt the new model is, no doubt, motivated as much by unfamiliarity and research inertia as it is by a careful evaluation of the merits of both versions. But there may be a different reason for this state of affairs, and one that may continue to prevail in the RST research. Some personality researchers appreciate that RST encapsulates some of the core elements of emotion and motivation, as they relate to personality, especially the focus on approach and avoidance as the two fundamental dimensions of behaviour. But they also think that the specific details of Gray's work are not entirely appropriate at the human level of analysis. For example, Carver and Scheier (1998; see Carver 2004) has made changes to the emotions

associated with reward and punishment systems. Their view of these systems are reflected in the broad-band BIS-BAS scales of Carver and White (1994), which may be seen as reflecting general motivational tendencies of avoidance and approach rather than the specifics of the BIS and BAS as detailed in Gray's work. This shows that a 'family' of RST-related theories has developed, which serves, depending on one's opinion, either to enrich or confuse the literature, especially when the same term ('BIS') is used to measure theoretically different constructs. Because the revised theory is even more specific about neural functions, derived largely from typical animal learning paradigms, there is little reason to think that this attitude will change once the revised theory is fully assimilated into RST thinking. In order to help researchers make a choice of hypotheses, this section details and contrasts the two versions of the theory.

Foundations of RST

Jeffrey Gray's approach to understanding the biological basis of personality followed a particular pattern: (a) first identify the fundamental properties of brain-behavioural systems that might be involved in the important sources of variation observed in human behaviour; and (b) then relate variations in these systems to existing measures of personality. Of critical importance in this two-stage process was the assumption that the variation observed in the functioning of these brain-behavioural systems comprises what we term 'personality' – in other words, personality does not stand apart from basic brain-behaviour systems, but rather is defined by them. As we shall see below, relating *a* to *b* has proved the major, and still unresolved, problem for RST.

Gray's work was also influenced by an appropriate respect for the implications of Darwinian evolution by natural selection. He took seriously the proposition that data obtained from (non-human) animals could be extrapolated to human animals (e.g., Gray 1987; see McNaughton and Corr, chapter 3). Gray's work may be seen in the larger scientific context foreshadowed by Darwin's (1859) prescient statement in the *Origin of Species*, 'In the distant future I see open fields for far more important researches. Psychology will be based on a new foundation, that of the necessary acquirement of each mental power and capability by gradation. Light will be thrown on the origin of man and his history.'

General theory of personality

Today, it may seem trite to link personality factors to emotion and motivational systems, but this neo-consensus did not prevail in the

1960s, when very few personality psychologists argued for the importance of basic systems of emotion underlying personality. It is a mark of achievement that Gray's (1970) hypothesis – novel as it was then in personality research – is today so widely endorsed. The emergence of a *neuroscience of personality* – an oxymoron not too long ago – was shaped in large measure by Gray's work. However, as we shall see below, the main elements of Gray's approach already existed in general psychology: like Hans Eysenck's (1957, 1967) theories, Gray's innovation was to put together the existing pieces of scientific jigsaw to provide the foundations of a general theory of personality. As with the construction of any complex structure, it is, indeed, prudent to have firm foundations – in the case of theory, verified concepts and processes from anywhere in the discipline (or from other disciplines) – upon which the further building blocks of theory may be placed. For this reason Gray, like Pavlov (1927) before him, advocated a twin-track approach: the conceptual nervous system (cns) and the *central nervous system* (CNS) (cf. Hebb 1955; see Gray 1972a); that is, the cns components of personality (e.g., learning theory; see Gray 1975) and the component brain systems underlying systematic variations in behaviour (*ex hypothesi*, personality). As noted by Gray (1972a), these two levels of explanation *must* be compatible, but given a state of imperfect knowledge it would be unwise to abandon one approach in favour of the other. Gray used the language of cybernetics, in the form of cns-CNS bridge, to show how the flow of information and control of outputs is achieved (e.g., the Gray-Smith 1969 Arousal-Decision model; see below). That RST focuses on a relatively small number of basic phenomena is in the nature of theory building; but this fact should not be interpreted, as it sometimes is, as implying that RST is restricted to explaining only these phenomena.

In contrast to Gray's general approach, Hans Eysenck adopted a very different 'top-down' one. His search for causal systems was determined by the structure of statistically-derived personality factors/dimensions. The possibility that the structure of these factors/dimensions may not correspond to the structure of causal influences was never seriously entertained. We shall have reason to question the premises underlying this particular assumption (see Corr and McNaughton, chapter 5). However, in one important respect, Eysenck's approach is viable: this was to understand the causal bases of *observed* personality structure, defined as a unitary whole (e.g., Extraversion and Neuroticism). For this very reason, it is perhaps not surprising to learn that Eysenck's causal systems never developed beyond the postulation of a small number of very general brain processes, principally the Ascending Reticular Activating Systems (ARAS), underlying the dimension of

introversion-extraversion and cortical arousal (for a summary, see Corr 2004). It should be noted that this was not a fault in Eysenck's work, because as argued elsewhere (Corr 2002a) there is considerable support for Eysenck's Extraversion-Arousal hypothesis and it does well to explain many forms of behaviours at the dimensional level of analysis. Taken together, Gray's and Eysenck's approaches are complementary, tackling important problems at different levels of analysis – we shall see below just how these levels of analysis can be integrated. Indeed, without Eysenck's work it is difficult to see how Gray's neuropsychological work would have led to a theory of *personality*. Also, Eysenck showed that a science of personality was possible and, in a wide variety of ways, of scientific importance (e.g., accounting for clinical neurosis).[1] (Fowles 2006 provides a superb summary of the development Gray's work.)

The 'Hull-Eysenck' and 'Mowrer-Gray' perspectives To understand the theoretical differences between the approaches adopted by Gray and Eysenck, it is necessary to delve into some of the scientific problems that dominated psychology during the middle of the twentieth century.

Eysenck's theory focused on a single factor underlying individual differences in arousal/arousability. This approach followed the well-trodden path of Hull (1952), whose learning theory concentrated on the single factor of drive reduction as underlying the effects of reinforcement. As noted by Gray (1975, p. 25), the 'Hullian concept of general drive, to the extent that it is viable, does not differ in any important respects from that of arousal'. To the extent that both Hull and Eysenck argued for one causal factor affecting learning, their position may be called the 'Hull-Eysenck perspective' (Corr, Pickering and Gray 1995a).

In contrast to this perspective – and reflecting the changes in learning theory that were taking place in general psychology – Gray's alternative position argued for a two-process theory of learning based upon reward and punishment systems. This position, dubbed the 'Mowrer-Gray perspective' (Corr *et al*. 1995a), reflected the importance of Mowrer's (1960) influential work in which he argued that learning is composed of two processes: (a) associative (Pavlovian) conditioning and (b) instrumental learning. In addition, and of particular significance for RST, Mowrer also argued that the effects of reward and punishment had different behavioural effects as well as different underlying bases.

[1] On a personal level, Gray was influenced by the fact that he undertook clinical and doctoral training in Eysenck's own Department, who encouraged him to translate Russian works on personality (see Corr and Perkins 2006).

Emotion was introduced in this learning account by Mowrer's theory that such states (e.g., hope) played the role of the internal motivator of behaviour (also see Konorski 1967; Mackintosh 1983). This two-factor (punishment/reward) theory was supported by neurophysiological findings; e.g., the discovery of the 'pleasure centres' in the brain in the 1950s (e.g., Olds and Milner 1954). Thus, from Mowrer's theory came the claim that (a) reward and punishment are different processes and (b) states of emotion serve as internal motivators of behaviour. To link this theory to individual differences in the functioning of brain-behavioural systems – a theoretical claim that also came out of Hull's work – and, then, to well-known personality factors was a logical step; although as obvious as it may now appear it takes a scientist of exceptional insight to recognize and appreciate its potential.

Standard (1982) RST

Eysenck's arousal theory of Extraversion (Eysenck 1967) postulated that introverts and extraverts differ with respect to the sensitivity of their cortical arousal system in consequence of differences in response thresholds of their Ascending Reticular Activating System (ARAS). According to this theory, compared with extraverts, introverts have lower response thresholds and thus higher cortical arousal. In general, introverts are more cortically aroused and more arousable when faced with sensory stimulation. However, the relationship between arousal-induction and actual arousal is subject to the moderating influence of transmarginal inhibition (TMI: a protective mechanism that breaks the link between increasing stimuli intensity and behaviour at high intensity levels): under low stimulation (e.g., quiet or placebo), introverts should be more aroused/arousable than extraverts, but under high stimulation (e.g., noise or caffeine), they should experience over-arousal which, with the evocation of TMI, can lead to lower increments in arousal as compared with extraverts; conversely, extraverts under low stimulation should show low arousal/arousability, but under high stimulation, they should show higher increments in arousal. A second dimension, Neuroticism (N), was related to activation of the limbic system and emotional instability (see Eysenck and Eysenck 1985). It was against this backdrop that RST developed.

Gray (1970, 1972b, 1981) proposed his alternative theory to Eysenck's. This theory proposed changes: (a) to the position of Extraversion (E) and Neuroticism (N) in factor space; and (b) to the neuropsychological bases of E and N. Gray argued that E and N should be rotated by approximately 30° to form the more causally efficient axes

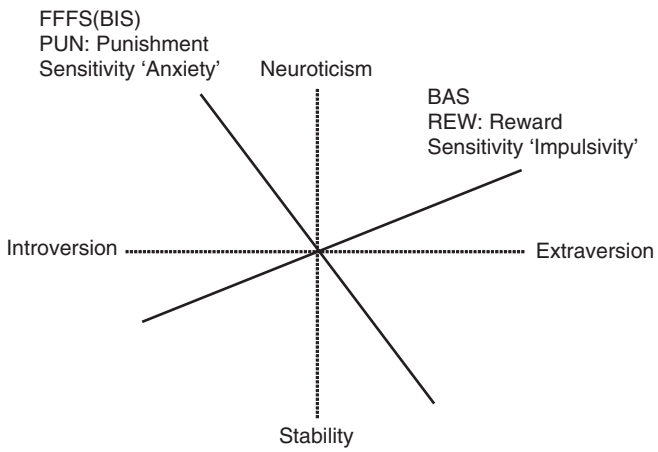

Figure 1.1 Position in factor space of the fundamental punishment sensitivity) and reward sensitivity (unbroken lines) and the emergent surface expressions of these sensitivities, i.e. Extraversion (E) and Neuroticism (N) (broken lines). In the revised theory, a clear distinction exists between fear (FFFS) and anxiety (BIS), and separate personality factors may relate to these systems; however, for the present exposition, these two systems are considered to reflect a common dimension of punishment sensitivity

of 'punishment sensitivity', reflecting Anxiety (Anx), and 'reward sensitivity', reflecting Impulsivity (Imp) (Figure 1.1; see Pickering, Corr and Gray 1999).[2]

In broad terms, the 1982 version of RST predicted that Imp+ individuals are most sensitive to *signals* of reward, relative to Imp– individuals; and Anx+ individuals are most sensitive to *signals* of punishment, relative to Anx– individuals. The orthogonality of the axes was interpreted to suggest: (a) that responses to reward should be the same at all levels of Anx; and (b) responses to punishment should be the same at all levels of Imp (this position has been named the 'separable subsystems hypothesis'; Corr 2001, 2002a). According to

[2] The relationship between Eysenck's and Gray's theories have not yet been fully clarified. For example, on the basis of empirical research, it seems likely that arousal is important in the initial conditioning of emotive stimuli which, then, serve as inputs into Gray's emotion systems; in turn, activation of these systems is expected to augment arousal and, thereby, influence conditioning processes quite independent of their role in generating emotion and motivational tendencies. If introversion-extraversion reflects the balance of reward and punishment sensitivities, then it may not be incompatible to argue that Eysenckian extraversion-arousal processes in conditioning continue to be relevant in Gray's RST.

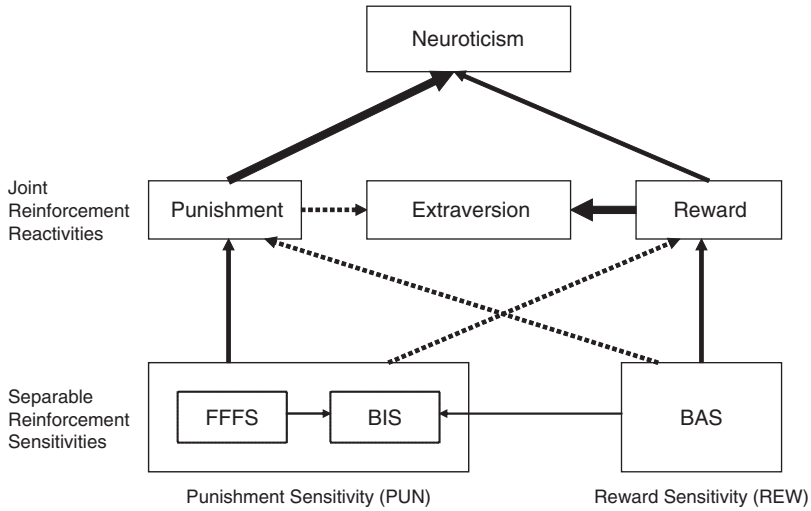

Figure 1.2 A schematic representation of the hypothesized relationship between (a) FFFS/BIS (punishment sensitivity; PUN) and BAS (reward sensitivity; REW); (b) their joint effects on reactions to punishment and reward; and (c) their relations to Extraversion (E) and Neuroticism (N). E is shown as the balance of punishment (PUN) and reward (REW) reactivities; N reflects their combined strengths. Inputs from the FFFS/BIS and BAS are excitatory (unbroken line) and inhibitory (broken line) – their respective influences are dependent on experimental factors (see text). The strength of inputs to E and N reflects the 30° rotation of PUN/REW and E/N: relatively strong (thick line) and weak (thin line) relations. The input from punishment reactivity to E is inhibitory (i.e. it reduces E), the input from reward reactivity is excitatory (i.e. it increases E). The BIS is activated by simultaneous activation of the FFFS and the BAS, and its activation increases punishment sensitivity. It is hypothesized that the joint effects of PUN and REW gives rise to the surface expression of E and N: PUN and REW represent the underlying biology; E and N represent their joint influences at the level of integrated behaviour

RST, Eysenck's E and N dimensions are derivative, secondary factors of these more fundamental punishment and reward sensitivities: E reflects the balance of punishment and reward sensitivities; N reflects their joint strengths (Gray 1981) (see Figure 1.2).

Gray's theory also explained Eysenck's arousal effects: *ex hypothesi*, on average, punishment is more arousing than reward, and introverts are more sensitive to punishment, therefore introverts experience more induction of arousal and tend to be more highly aroused. In contrast,

Eysenck maintained that, to the extent that reinforcement effects are mediated by personality, they are a consequence of arousal level and not sensitivity to reward and punishment per se.

Clinical neurosis

According to Eysenck's arousal theory, introverts are prone to suffer from anxiety disorders because they more easily develop classically conditioned (emotional) responses; this theory was expanded with the inclusion of incubation effects in conditioning effects (Eysenck 1979) to account for the 'neurotic paradox' (i.e., the failure of extinction with continued non-reinforcement of the conditioned stimulus (CS)); coupled with emotional instability, reflected in N, this made the introverted neurotic (E–/N+) especially prone to the anxiety disorders.

However, from the inception of this arousal-based theory of personality, there were a number of problems. First, introverts show *weaker* classical conditioning under conditions conducive to high arousal (e.g., in eye-blink conditioning; Eysenck and Levey 1972); and a crossover pattern of E × arousal is easily confirmed (e.g., in procedural learning; Corr, Pickering and Gray 1995b), supporting Eysenck's *own* theory that introverts are transmarginally inhibited by high arousal (see above). Other problems attend Eysenck's arousal-conditioning claims. For example, Imp (inclined into the N plane), not sociability, is often associated with conditioning effects (Eysenck and Levey 1972); this should place high arousability, and thus high conditionability, in the stable-introverted quadrant defined by E×N space, not in the neurotic-introvert quadrant required by the theory and clinical data. Thus, Eysenck's theory seems unable to explain the aetiology of anxiety in neurotic-introverts, which was one of the major points of the theory from its early days. Time of day effects further undermine the central postulates of Eysenck's personality theory of clinical neurosis. Gray (1981) provides a masterly discussion of these problems, which according to him thrusts a dagger into the heart of Eysenckian theory.

Conditioning and emotion Gray identified a more compelling reason for rejecting the classical conditioning theory of neurosis. In classical conditioning theory, as a result of the conditioned stimulus (CS) and unconditioned stimulus (UCS) being systematically paired, the CS comes to take on many of the eliciting properties of the UCS: when presented alone, the CS produces a response (i.e., the conditioned response (CR)) that resembles the unconditioned response (UCR) elicited by the UCS. Thus innate fear (UCS) may be elicited by a CS: hence the

classic conditioning idea of neurosis. As so often the case, the devil is in the detail. The problem is that the CR does not substitute for the UCR – in several important respects, the CR does not even resemble the UCR. For example, a pain UCS will elicit a wide variety of reactions (e.g., vocalization and behavioural excitement) which are quite different to those elicited by a CS *signalling* pain: the latter produces anxiety and a different set of behaviours (e.g., quietness and behavioural inhibition). Thus, classical conditioning cannot explain the pathogenesis or phenomenology of neurosis, although it can explain how initially neutral stimuli (CSs) acquire the motivational power to elicit this state. Well, if the CR is not simply a version of the UCR then what generates the negative emotional state that characterizes neurosis? Gray's claim was an innate mechanism, namely the *Behavioural Inhibition System* (BIS) (Gray 1976, 1982).

Three systems of standard RST

RST gradually developed over the years to include three major systems of emotion:

(1) The Fight-Flight System (FFS) was hypothesized to be sensitive to *unconditioned* aversive stimuli (i.e., innately painful stimuli), mediating the emotions of rage and panic – this system was related to the state of negative affect (NA) (associated with pain) and Eysenck's trait of Psychoticism.
(2) The Behavioural Approach System (BAS) was hypothesized to be sensitive to *conditioned* appetitive stimuli, forming a positive feedback loop, activated by the presentation of stimuli associated with reward and the termination/omission of signals of punishment – this system was related to the state of positive affect (PA) and the trait of Imp.
(3) The Behavioural Inhibition System (BIS) was hypothesized to be sensitive to *conditioned* aversive stimuli (i.e., signals of both punishment and the omission/termination of reward) relating to Anx, but also to extreme novelty, high intensity stimuli, and innate fear stimuli (e.g. snakes, blood) which are more related to fear.

With respect to the CNS, Gray used data from a wide range of sources, principally (a) the effects of lesion of specific neural sites on behaviour and (b) the effects of drugs – initially the barbiturates and alcohol, and later anxiolytics – on specific classes of behaviour. Gray's 'philosopher's stone' was the detailed pattern of behavioural effects of classes of drugs known to affect emotion in human beings; in this way anxiety could be

operationally defined as those behaviours changed by anxiolytic drugs. The obvious danger of circularity of argument was avoided by the postulation that anxiolytic drugs do not simply reduce anxiety (itself a vacuous tautology), but could be shown to have a number of behavioural effects in typical animal learning paradigms. It turned out that such drugs affected conditioned aversive stimuli, the omission of expected reward and conditioned frustration, all of which acted on a postulated Behavioural Inhibition System which was charged with the task of suppressing ongoing operant behaviour in the face of threat and enhancing information processing. Later, the Behavioural Approach System was added to account for behavioural reactions to rewarding stimuli, which was largely unaffected by anxiety-reducing drugs. The circularity of this argument was further broken by the behavioural profile of the newer classes of anxiolytics which, as it turned out, had the same behavioural effects, and acted on the same neural systems, as the older class of drugs, despite the fact that they had different psychopharmacological modes of action and side-effects (Gray and McNaughton 2000).

Revised (2000) RST

Chapters 2 and 5 provide a detailed account of the neuropsychology of the Gray and McNaughton (2000) revised theory. This section provides a brief overview of this new theory, which shows that there are a number of significant changes to the systems that hold important implications for conceptualization and measurement.

Revised RST postulates three systems.

(1) The Fight–Flight–Freeze System (FFFS) is responsible for mediating reactions to *all* aversive stimuli, conditioned and unconditioned. A hierarchical array of modules comprises the FFFS, responsible for avoidance and escape behaviours. Importantly, the FFFS mediates the 'get me out of this place' emotion of fear, not anxiety. The FFFS is an example of a negative feedback system, designed to reduce the discrepancy between the immediate threat and the desired state (i.e., safety). The associated personality factor comprises fear-proneness and avoidance, which clinically mapped onto such disorders as phobia and panic. (In contrast, the original, 1982, theory assigned the FFFS to reactions to *unconditioned* aversive (pain) stimuli.)

(2) The Behavioural Approach System (BAS) mediates reactions to *all* appetitive stimuli, conditioned and unconditioned. This generates the appetitively hopeful emotion of 'anticipatory pleasure'. The associated personality comprises optimism, reward-orientation and impulsiveness, which clinically maps onto addictive behaviours (e.g., pathological

gambling) and various varieties of high-risk, impulsive behaviour, and possibly the appetitive component of mania. (The BAS is largely unchanged in the revised version of RST.) This is a positive feedback system, designed to move away from current appetitive goal-state towards the biological reinforcer.

It is important to distinguish the incentive motivation component and the consummatory component of the reactions to unconditioned appetitive stimuli. Gray believed that no single system mediates the consummatory component of such reactions (e.g., reactions to unconditioned stimuli): e.g., copulation and eating/drinking involve very different response systems (also see below for further work needed to the concept of the BAS). The BAS is responsible for reducing the temporo-spatial distance between current appetitive goal state and the final biological reinforcer.

(3) The Behavioural Inhibition System (BIS) is now responsible, not for mediating reactions to conditioned aversive stimuli and the special class of innate fear stimuli, but for the resolution of goal conflict in general (e.g., between BAS-approach and FFFS-avoidance, as in foraging situations, but it is also involved in BAS-BAS and FFFS-FFFS conflict). It is a negative feedback system aimed at countering the deviation from the reference state of no goal conflict. The BIS generates the 'watch out for danger' emotion of anxiety, which entails the inhibition of prepotent conflicting behaviours, the engagement of risk assessment processes, and the scanning of memory and the environment to help resolve concurrent goal conflict. The BIS resolves conflicts by increasing, by recursive loops, the negative valence of stimuli (these are adequate inputs into the FFFS), until behavioural resolution occurs in favour of approach or avoidance. Subjectively, this state is experienced as worry and rumination. The associated personality comprises worry-proneness and anxious rumination, leading to being constantly on the look-out for possible signs of danger, which clinically maps onto such conditions as generalized anxiety and obsessional-compulsive disorder (OCD) – both conditions reflect a lack of adequate goal conflict resolution appropriate to local environmental parameters. There is an optimal level of BIS activation: too little leads to risk-proneness (e.g., psychopathy) and too much to risk aversion (generalized anxiety), both with sub-optimal conflict resolution. It is intriguing to speculate that modern-day angst, and social malaise, is in part due to the conflict induced by reward-reward conflicts (e.g., which holiday to go on, which car to purchase and which career to pursue): choice per se has a negative component. The way in which the FFFS, BIS and BAS relate to Extraversion and Neuroticism is shown in Figure 1.2.

The revised Gray and McNaughton (2000) theory makes a number of new claims, some of which contradict those derived from unrevised RST theory.

(1) In contrast to the 1982 theory, the distinction between the BIS and FFFS (defensive direction) is totally divorced from the conditioned or unconditioned nature of stimuli used to elicit emotion. In the 1982 version of the theory, the BIS was activated by conditioned aversive stimuli (as well as 'innate fear stimuli'), and the Fight-Flight System (as it was then called) by unconditioned aversive stimuli. This conditioned-unconditioned distinction is now not relevant to the revised theory: both types of stimuli can activate the FFFS and, provided there is conflict, the BIS. The importance of the conditioned-unconditioned distinction seems to have come from the strong correlation between, on the one hand, unconditioned and immediate threat, and, on the other hand, conditioned and potential threat. It also turned out that many, but not all, forms of conditioned stimuli are, in fact, conflict stimuli, and BIS effects were typically measured as the suppression of ongoing BAS-controlled behaviour (e.g., conditioned emotional suppression).

(2) There is now a sharp (ethological, behavioural and pharmacological) distinction between fear (FFFS) and anxiety (BIS). This distinction is still controversial. Consistent with the theory, Barlow (1988) associates anxiety with future danger and fear with imminent danger, but others challenge this distinction (see Fowles 2000). However, the theory shows how these two emotions are different. This difference is based on the concept of 'defensive direction': fear refers to the elicitation of a range of reactions that have the common function of facilitating the movement of the animal *away from* threat; anxiety refers to the elicitation of a range of reactions that have the common function of facilitating the movement of the animal *towards* threat (or more generally resolving conflict). The concept of defensive direction provides a single principle to define inputs to the BIS – the 1982 theory provided an essentially ad hoc list.

This conceptualization is based on the Blanchards' (e.g., Blanchard and Blanchard 1990) etho-experimental work which linked to a state of fear a set of behaviours elicited by a predator. These behaviours turn out to be sensitive to drugs that are panciolytic (i.e., panic reducing), but not to those drugs that are only anxiolytic (i.e., anxiety-reducing). Such behaviours include simple avoidance (fleeing), freezing and defensive attack.[3] In contrast, they link to a state of anxiety a quite different set of

[3] As noted by Eilam (2005), in fleeing, the prey physically removes itself from the vicinity of the predator; in freezing, the prey remains immobile in order to evade the attention of the predator; and in fighting (or defensive threat), the prey heads towards the predator in

behaviours (especially 'risk assessment'), elicited by the potential presence of a predator that turn out to be sensitive to anxiolytic drugs. Because of the detailed effects of anxiolytic drugs on behaviour (see Gray and McNaughton 2000), it is argued that the key factor distinguishing fear and anxiety is not that posited by the Blanchards, namely immediacy (or certainty) versus potentiality (or uncertainty) of threat but 'defensive direction': fear operates when *leaving* a dangerous situation (active avoidance), anxiety when *entering* it (e.g., cautious 'risk assessment' approach behaviour) or withholding entrance (passive avoidance).

(3) An important feature of the revised theory is it now explains the phenomenology of fear and anxiety in Darwinian adaptive terms; and at the specific neural module levels, specific reactions serve particular adaptive functions. Few theories of fear and anxiety attempt to explain *why* these emotions have their specific natures: *why* should anxiety should be related to rumination, worry, risk assessment, vigilance for potentially bad things?

(4) There are distinct systems in the brain that control specific functional classes of behaviour (e.g., fight, flight, freezing as separate classes). These systems can be viewed as the targets of particular perceptions/cognitions ('I am about to be eaten by the cat/lion' for rat/human respectively). They can also be viewed as the sources of particular emotional behaviours (e.g., panic) that, if excessive or inappropriate, represent particular types of clinical symptom (e.g., panic attack) or syndrome (e.g., panic disorder). These local systems are organized into clusters that control more global functional classes of behaviour (e.g., defence). Where several specific classes of behaviour (fight, flight, freeze) all have a high probability of being elicited in essentially the same global situation (predatory threat) their organization (through the course of evolution) into a more global system allows co-ordination and selection of just one of the primed classes of behaviour. Modulatory systems can affect specific global functional classes (e.g., threat sensitivity) or many together (e.g., arousal, attention).

(5) All of the above levels of neural organization can be assigned both state ('How active are they right now'; see chapter 2) and trait ('How reactive are they in general to a fixed stimulus'; see chapter 5). Macroscopic factors, affecting global classes of functional system (e.g., defence) or cutting across such systems (e.g., arousal), contribute to personality. Personality can be seen as reflecting global functional variations in these systems.

order to discourage its predatory behaviour – defensive fighting occurs when the prey has no possibility of freezing and fleeing and must face the predator.

(6) An important point for the personality theorist is the layer of complexity in the old (Gray and Smith 1969) and new theories – however, this is usually ignored in research – that focuses on the parametric interactions between approach and avoidance systems when each is concurrently activated. The key point is that when the BAS and FFFS are activated unequally (i.e., when there is little conflict between approach and avoidance), they nonetheless interact. This interaction is symmetrical. Activation of one system inhibits the other. This interaction (in its purest form counterconditioning of one stimulus by a motivationally opposite stimulus) is insensitive to anxiolytic drugs and so practically as well as theoretically independent of the BIS. Thus, while the two systems are independent in that changes in the sensitivity of one will not affect the sensitivity of the other, they are not independent in that concurrent activation will cause interactions in their generation of behaviour output. The primary symmetrical interactions between the systems are also non-linear, accounting for such phenomena as behavioural contrast and peak shift (Gray and Smith 1969). Joint activation increases arousal while producing a subtractive effect on decision (Corr and McNaughton, chapter 5). This important matter is considered in detail below. Superimposed on these symmetrical interactions is the BIS. This is activated more as the difficulty of resolving the decision between the two (approach-avoid) increases (i.e., as the relative power of approach and avoidance become more equal). Its activation results in asymmetrical effects. It boosts arousal (over and above the additive effect of the existing conflicting motivations) while it amplifies activity in aversive system but not the appetitive one. Under conditions of conflict, then, it increases risk aversion.

(7) Abnormal levels of expressions of personality may result from three conditions: (a) as a normally adaptive reaction to their specific eliciting stimuli (e.g., mild anxiety before important examination); (b) at maladaptive intensity, as a result of excessive sensitivity to their specific eliciting stimuli (e.g., sight of harmless spider = fearful avoidance); and (c) at maladaptive intensity, as a result of excessive activation of a related structure by its specific eliciting stimuli but where the 'symptoms' are not excessive given the level of input (e.g., oncoming train = panic).

Defensive distance

Revised RST contends that defensive behaviour results from the superimposition on defensive direction (i.e., approach or avoid) of what is known as 'defensive distance'. According to this two-dimensional

Table 1.1 *Relationship between actual and perceived defensive distance in low, medium and high fearful individuals*

System state	Defensive distance	Real distance sufficient to elicit reaction
Low defensive individual:	Perceived distance > actual distance	Short
Normal defensive individual:	Percieved distance = actual distance	Medium
High defensive individual:	Perceived distance < actual distance	Long

model of McNaughton and Corr (2004), for a particular individual in a particular situation, defensive distance equates with real distance; but, in a more dangerous situation, the perceived defensive distance is shortened. In other words, defensive behaviour (e.g., active avoidance) will be elicited at a longer (objective) distance with a highly dangerous stimulus (corresponding to shortened perceived distance), as compared to the same behaviour with a less dangerous stimulus. According to the theory, neurotic individuals have a much shorter perceived defensive distance, and thus react more intensively to relatively innocuous (real distance) stimuli. For this reason, weak aversive stimuli are sufficient to trigger a neurotic reaction in highly defensive individuals; but for the braver individual, aversive stimuli would need to be much closer to elicit a comparable reaction. This set of relations is shown in Table 1.1 (taken from Corr and Perkins 2006).

Defensive distance operationalizes an internal cognitive construct of intensity of perceived threat. It is a dimension controlling the type of defensive behaviour observed. In the case of defensive avoidance, the smallest defensive distances result in explosive attack, intermediate defensive distances result in freezing and flight, and very great defensive distances result in normal non-defensive behaviour. The notion that there is a 'Distance-Dependent Defence Hierarchy' goes back a long way (e.g., Ranter 1977). Defensive distance maps to different levels of the FFFS (see chapter 2): as an animal cannot freeze and flee at the same moment, these behaviours must be controlled by different bio-behavioural mechanisms (Eilam 2005), and here we see the importance of a 'hierarchical' arrangement of defensive models, with higher-level modules inhibiting lower-level ones. The psychological state experienced at very short defensive distance would be labelled panic, which is

often associated, at least at clinical extremes, with the cognition 'I'm going to die' – we may liken this cognition to whatever cognition runs through the rat's mind when nose-to-nose with a dangerous predator (e.g., hungry cat). The rat's cognition emotion may be similar to the emotion we would feel if trapped in a car in the path of an oncoming high-speed train. At intermediate defensive distances, we would probably substitute panic for phobic avoidance (e.g., not driving over the train line if there is any chance of an oncoming train). With the opposite direction, defensive quiescence occurs; at intermediate distances, risk assessment is observed; and, at very long distances, defensive behaviour fades to be replaced by normal pre-threat behaviour. The first experimental test of this theory in human beings was made by Perkins and Corr (2006), who presented threat scenarios and correlated known measures of personality (including fear and anxiety scales) with the intensity and direction dimensions of chosen responses. The results broadly supported the differentiation of fear-related *avoidance* of threat and anxiety-related *cautious approach* to threat.

McNaughton and Corr (2004) view individual differences in defensive distance for a fixed real distance as a reflection of the personality dimension underlying 'punishment sensitivity', or 'threat perception' (or neurotic-introversion), which affects the FFFS directly, and the BIS indirectly (e.g., via FFFS-BAS goal conflict). Anxiolytic drugs alter (internally perceived) defensive distance relative to actual external threat. They *do not* affect defensive behaviour directly, but rather shift behaviour along the defensive axis, often leading to the output of a different behaviour (e.g., from freezing to flight; for discussion of a fundamental part of revised RST, see chapter 2). The modulation acts like a magnification factor. An important corollary of this claim is that comparison of individuals on a single measure of performance at only a single level of threat may produce confusing results (e.g., one person may be in a state of panic and so cease moving; another may actively avoid and so increase their movement). In other words, highly sensitive and insensitive fearful individuals will show *different* behaviours *at the same level of threat* (defined in objective terms), as indeed will trait-identical individuals at different levels of threat. Thus, moving people along this axis of defensive distance (by drugs or by experimental means) will not simply affect the strength or probability of a given behaviour, but is expected to result in different behaviours. Thus, at the core of the revised theory are ethological considerations: specific behaviours relate to specific threats and environmental conditions.

A philosophical digress: utility theory

We have now covered the central tenets of the old and new versions of RST; the state and trait aspects of the theory are expanded in chapters 2 and 5, respectively. At this point, it might be appropriate to stand back from the details of the theory and consider RST in the broader domain of philosophy. Although not formally part of RST it is perhaps worth noting that the central tenets of RST, i.e., that behaviour is governed by two major affective dimensions of pleasure and pain finds an echo in the view of the English philosopher, Jeremy Bentham (1748–1832), who formulated Utilitarian theory, which argues that society (and government public policy) should follow the principle of the 'greatest happiness to the greatest number'. Bentham's philosophy arose out of his views on the nature of individual behaviour; he wrote in *Introduction to the Principles of Morals and Legislation* (1781):

> Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do as well as to determine what we shall do. On the other hand, the standard of right and wrong, on the other chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think; every effort we can make to throw off our subjection, will serve but to demonstrate and confirm it. In words a man may pretend to abjure their empire: but in reality he will remain subject to it all the while.

Bentham introduced the principle of the individual's 'hedonic calculation', which maximizes the utility of the individual; in other words, individuals seek to maximize their 'happiness' (defined as the surplus of pleasure over pain), or minimize their pain (defined as the surplus of pain over pleasure). As a philosophy of individual behaviour, Bentham's view may be seen to border on the obvious and circular; but this circularity is broken once we have a scientific theory of the hedonic calculus – and this is what RST offers.

This view of the governance of individual behaviour may be likened to the process of natural selection, that is, according to Darwin (1859/1968, p. 859) 'scrutinising, throughout the world, every variation, even the slightest; rejecting that which is bad ["pain"], preserving and adding up all that is good ["pleasure"]'. It is trite, but true, to say that most animals seek to minimize pain and maximize pleasure (although, in the case of the human animal, not necessarily in the crass, overt form that the word 'pleasure' usually implies).

In his consideration of 'individual hedonic calculus', Bentham noted that the wealthier a person is, the greater their total happiness – all else being equal, it is better to be rich than to be poor! However, the

wealth-happiness function is not linear; there is a 'marginal utility' function at work which states that the greater amount of utility a person *already* has (i.e., 'happiness'), the smaller will be the utility associated with an extra increment in wealth. (This principle of marginal utility is central to economic thinking today and is the rationale for, among other things, progressive rates of taxation.) Couched in the prosaic nomenclature of RST, this marginal utility function is likely to produce a non-linear relationship between an extra unit of reward (defined in some experimental manner) and experienced 'pleasure' (i.e., BAS-related emotion and behaviour). Perhaps it is for this reason that highly reward-sensitive gamblers need a large increment in reward to experience a perceptible increment in 'pleasure' – this general principle should hold also for highly BAS active individuals.

Bentham went on to state that the individual's hedonic calculus values pleasure or pain according to a number of parameters: (a) duration, (b) intensity, (c) certainty (or probability) and (d) propinquity vs. remoteness. It turns out that these are some of the very criteria that are important in operationalizing experimental variables in the RST laboratory. For example, certainty relates to defensive direction: a certain threat, of sufficient intensity and duration, should be avoided; propinquity vs. remoteness relates to defensive distance.

What this detour into political philosophy shows is that the types of constructs contained in RST do have a long tradition in intellectual thought and continue to underpin many public, and private, policies. Consideration of these domains lend support to the claim that RST, whilst perhaps not providing a complete account of emotion, behaviour and personality, focuses on some of their fundamental processes, as revealed by scientific findings as well as the wider realms of philosophical thought.

Future

We have now surveyed the main elements of RST, as encapsulated in their 1982 and 2000 versions. In this section we turn our attention to the future, asking which elements of RST require further development.[4]

The BAS and its parts

The Gray and McNaughton (2000) theory has little to say that is new about the BAS. But since the early 1990s, there has been debate in the

[4] The content of this section was influenced by discussions with two RST researchers, Mr Adam Perkins and Dr Andrew Cooper.

literature concerning its structure and psychometric properties. This section aims to show that the BAS is more complex than often thought, and that this complexity is not restricted only to its psychometric delineation.

Evolution of BAS complexity

There are several reasons for assuming that the BAS is much more complex than the FFFS (which motivates simple avoidance/escape), or indeed the BIS (which has a relatively simple process to resolve goal conflict). From an evolutionary perspective, this complexity may derive from the 'arms race' between predator and prey. The 'Life-Dinner Principle' (Dawkins and Krebs 1979) suggests that the evolutionary selective pressure on prey is much stronger than on predators: if a predator fails to kill its prey then is has lost its dinner, but if the prey fails to avoid/escape being the predator's dinner then it has lost its life. Although defensive behaviour, principally freezing, fleeing and defensive attack, are themselves relatively complex (Eilam 2005), it is nonetheless true that the behaviour of prey is intrinsically simpler than that of predator: all it has to do is avoid/escape – it really is life-or-death behaviour. In contrast, the predator has to develop counter-strategies to meet its BAS aims (to get its dinner, etc.), which entail a higher degree of cognitive and behavioural sophistication. This Life-Dinner Principle is related to the second reason for complexity, namely, heterogeneity of appetitive goals (e.g., securing food and finding/keeping a sexual mate), which demand a heterogeneity of BAS-related strategies. No one set of behaviours would be sufficient to achieve these very different BAS goals; therefore, it seems essential that the BAS entails a much more flexible repertoire of behaviours and planning processing than either the FFFS or the BIS.

BAS functions

The *primary* function of the BAS is to move the animal up the temporo-spatial gradient to the final biological reinforcer – it is for this reason that we should prefer the term 'approach' to the less precise 'activation'. This primary function is supported by a number of *secondary* processes. In its simplest form, the secondary process could comprise simple approach, perhaps with BIS activation exerting behavioural caution at critical points, designed to reduce the distance between current and desired appetitive state (e.g., as seen in foraging behaviour in a densely vegetated field); but in the case of human behaviour, this depiction of BAS-controlled approach behaviour is grossly oversimplified.

First, it is necessary to distinguish the *incentive* motivation component and the *consummatory* component of reactions to appetitive stimuli, as suggested by their distinct neuroanatomical substrates (Carver 2005). The neural machinery controlling reactions to unconditioned (innate) stimuli, and its associated emotion, must be different from that controlling the behaviour and emotion associated with *approach*, signalled by conditioned stimuli, to such stimuli. Even in the Gray and McNaughton (2000) revised theory, the BAS is still not sensitive to unconditioned stimuli.[5]

Second, moving to approach proper, we can discern a number of relatively separate, albeit overlapping, processes. At the simplest level, there seems an obvious difference between the 'interest' and 'drive' that characterizes the early stages of approach, and the behavioural and emotional excitement as the animal reaches the final biological reinforcer. Emotion in the former case may be termed 'anticipatory pleasure' (or 'hope'); in the latter case something akin to an 'excitement attack' – the resemblance with 'panic attack' is deliberate.[6]

There is, indeed, evidence, at the psychometric level of analysis, that the BAS behaviour/emotion is multidimensional. For example, the Carver and White (1994) BIS/BAS scales measure three aspects of BAS: reward responsiveness, drive and fun-seeking. As noted by Carver (2005, p. 9):

> The three aspects of BAS sensitivity that are reflected in the three BAS scales derive from theoretical statements about the ways in which BAS functioning should be reflected experientially. That is, high BAS sensitivity should cause people to seek new incentives [reward responsiveness], to be persistent in pursuing incentives [drive], and to respond with positive feelings when incentives are attained [fun seeking].

In the conceptualization favoured here, Drive is concerned with actively pursing desired goals, and reward responsiveness is concerned with excitement at doing things well and winning, especially to rewarding

[5] However, in Gray and McNaughton (2000), Fig. 5.1 (p. 86) incorrectly shows adequate inputs to the BAS to include unconditioned reward (Gray, personal communication). To emphasize the important distinction between reactions to unconditioned and conditioned appetitive stimuli, Gray (personal communication) wryly commented: 'Try copulating with a ham sandwich!'

[6] 'Excitement attacks' may not only occur at the consummatory stage of approach – in a similar way panic attacks do not only occur when life is threatened – but may be triggered at the conclusion of fulfilment of important sub-goals in the cascade of approach behaviour. Indeed, such 'highs' would seem essential to maintain motivation directed to final-goal directed behaviour when approach entails a series of sub-goal procedures. (I am indebted to Margaret Wilson who, in sharing the experience of her own excitement attacks, first brought this felicitous term to my attention.)

stimuli associated with fulfilling sub-goal procedures: both processes seem to reflect the process of behavioural maintenance needed during complex approach behaviour involving multiple sub-goals. In contrast, fun-seeking may relate more to behaviours closer to the final biological reinforcer, which no longer entails planning and restraint of behaviour – fun-seeking is similar to impulsivity in this respect (see below). It is unlikely that *these* specific traits adequately capture the true nature of BAS behaviour, but they do usefully measure relatively separate (but overlapping) processes.

Sub-goal scaffolding

In order to move along the temporo-spatial gradient to the final primary biological reinforcer, it is necessary (at least in human beings) to engage in sub-goal scaffolding. This process consists of (a) identifying the biological reinforcer, (b) planning behaviour, and (c) executing the plan (i.e., 'problem solving') at each stage of the temporo-spatial gradient – this is in accordance with the type of cognitive operations first discussed by Miller, Gelenter and Pribram (1960), *Plans and the Structure of Behavior*.[7] Now, complex approach behaviour entails a series of behavioural processes, some of which oppose each other. For example, behaviour *restraint* and *planning* are often demanded to achieve BAS goals, but not at the final point of *capture* of the biological reinforcer, where non-planning and fast reactions (i.e., impulsivity) are more appropriate. Just being impulsive – that is, acting fast without thinking and not planning – would lead to being stranded on 'local highs' (in formal problems solving terms), moving the animal along the temporo-spatial gradient *away* from the final biological reinforcer. For this reason 'impulsivity' is perhaps not the most appropriate term for the personality factor corresponding to the full range of processes entailed by the BAS.

Another way to look at restraint and impulsiveness in BAS approach is to think of closed and open feedback systems. A closed feedback system entails feedback which modifies behaviour. In the case of BAS this may entail some degree of restraint. But in an open feedback system there is no feedback to affect perception and behaviour: the output simply executes on the assumption that the consequence will be as intended (Carver and Scheier 1998). The latter form of feedback system is

[7] These authors argued that behaviour is guided by plans and goals and (self) regulated by discrepancy reducing feedback processes. They also noted that any general goal can be broken down into sub-goals; but this raises the problem of the control of sub-goals, which usually demands some form of hierarchical system of control of action plans.

appropriate to reflex-like responses, where there is little time for feedback to be processed. In the case of dysfunctional impulsivity (cf. Dickman 1990), it seems that the open feedback system is triggered long before environmental conditions warrant, with the result that behaviour comprises non-planning, lack of reflection and rigid behavioural repertoire (cf. Patterson and Newman 1993).

Sub-goal scaffolding, which is necessary for planning effective BAS approach to appetitive stimuli, will often entail the *inhibition* of impulsive behaviour, and for this reason we might suspect that BAS behaviours are hierarchically organized such that lower-level reactions (e.g., impulsiveness) are inhibited by high-level (control) modules, which involve the cognitive processing underlying sub-goal scaffolding.[8] In parallel with the example of FFFS-mediated panic attack, having an impulsivity-related behaviour when the biological reinforcer (i.e., unconditioned stimulus) is not proximal would be inappropriate. A panic attack is appropriate when suffocating; rash impulsivity is appropriate when cognitive planning can be replaced, at short temporo-spatial distance, by fast 'getting', or a physical grabbing, action (Carver 2005). Therefore, there is a need to take due consideration of two processes in BAS-controlled approach: (a) *behavioural restraint* is needed to plan and execute effective sub-goal scaffolding; and (b) *impulsive behaviour* is needed to get/capture the final biological reinforcer at near-zero temporo-spatial distance. However, this is not to imply that the emotional component of BAS behaviour would be attenuated at the early stages of approach behaviour; in fact, as noted above, the fulfilment of sub-goals is likely to entail periodic bursts of emotional excitement.

This restraint-impulsivity dimension, which is argued here to co-vary along a dimension of temporo-spatial distance to goal, may be illustrated by reference to the behaviour of careful financial planners and pathological gamblers. If the goal is to accumulate wealth, then gambling is an inappropriate strategy. In order to achieve this goal, restraint of impulsivity is needed, and short-term gains must be sacrificed for long-term success. As noted by Carver (2005, p. 312), 'unfettered impulse can interfere with the attainment of longer term goals'. This process may be labelled 'temporal bridging' to emphasize the need to maintain approach behaviour across time gaps during which approach behaviour

[8] It is to be expected that where goal conflict is present then the BIS will be engaged. This BIS influence on BAS functions provides another juncture at which systems interact. Such interaction may give rise to important effects on the BAS. For example, a hyperactive BIS would significantly disrupt functioning of the BAS by producing too much hesitation and risk assessment, thus impairing the adaptive approach behaviour.

is not being immediately reinforced. It would not make sense to define high BAS activity in terms of reckless impulsive behaviour that fails to achieve BAS ends: a longer timeframe is required to see fully how planning, involving scaffolding of sub-goals and temporal bridging, serves BAS ends.

Impulsivity The concept of *sub-goal scaffolding* may shed new light on the role played by the trait of impulsivity in BAS behaviours. We have already seen that there is often the need for considerable planning in BAS behaviour, including reflection on likely outcome of alternative courses of action, for the BAS to achieve its goals. Let us now consider a typical measure of impulsivity (I_7; Eysenck, Pearson, Easting and Allsopp 1985), which is defined by aspects of fast reactions and non-planning:

(1) Do you often buy things on impulse?
(2) Do you often do things on the spur of the moment?
(3) Do you generally do and say things without stopping to think?
(4) Do you often get into a jam because you do things without thinking?

Such behaviours are insufficient to account for the full range of BAS-related processes and behaviours required to achieve BAS objectives. Differentiating functional and dysfunctional forms of impulsivity (Dickman 1990) does not help to resolve this debate. According to the position advanced here, 'dysfunctional' impulsivity is nothing more than the impulsive behaviour displayed at an inappropriate stage in the series of BAS processes involved in approach (e.g., as in the above example, pursuing the goal of being wealthy by engaging in impulsive gambling behaviour). There is a further conceptual confusion resulting from relating impulsivity to the BAS. This problem arises because impulsive behaviour may arise from either (a) an underactive BIS or (b) an overactive BAS (Avila 2001) (see Avila and Torrubia, chapter 7). As argued in chapter 5, impulsivity as a high-level personality construct may reflect the functioning of several underlying systems and not simply one (e.g., the BAS).

Item response theory

As discussed in chapters 5 and 6, there have been many attempts to develop psychometric measures of RST constructs: the aforementioned Carver and White BIS/BAS scales have been the most popular. The data pertaining to how these scales relate to the experimental manipulation of

RST variables has, however, proved problematic. The reason for this state of the literature may have much to do with inadequate psychometric definition of the central constructs of RST, as discussed above in the case of the BAS. However, there is a second factor that needs to be considered: the *precision* of measurement across the whole range of reinforcement sensitivities.

As argued by Gomez, Cooper and Gomez (2005), it is highly desirable to apply *Item Response Theory* (IRT) to ensure that RST scales contain sufficient items to measure reinforcement sensitivities along the entire length of the latent trait. Gomez *et al*. (2005) used IRT to examine the psychometric properties of the BIS/BAS scales and found that, although all items in all four scales (one BIS, three BAS) were reasonably effective in measuring their scales' designations, their precision of measurement was adequate only for low to moderately high trait level range: high sensitivities in particular are poorly measured, and it is in these extreme groups that RST has most interest. An item bank of FFFS, BAS and BIS items for all levels of the latent traits would prove a valuable addition to RST research. Such a bank of items, maybe using computerized adaptive testing to present the items relevant to the participant's latent trait, is needed to test with adequate precision in different populations experimental predictions based on RST constructs.

Basic and complex emotions

At its present stage of development, RST does not provide a complete account of emotion processing; rather it has focused on two fundamental negative emotions systems (underlying fear and anxiety),[9] and one positive emotion system (underlying appetitive drive and anticipatory pleasure). It has not addressed basic emotions (e.g., disgust and sadness). (Gray was working on the neural basis of disgust before his death, and had already published on this topic; e.g., Phillips *et al*. 2004.)

Gray (1985, 1994) suggested that emotional states are the blending of the more basic FFFS, BIS and BAS; for example, sadness may result from being confronted by punishment, which has to be approached, but

[9] It is interesting to note that the emotion of 'anxiety' does not feature in Ekman's list of basic emotions, and nor does an 'approach' emotion, perhaps because of restricted and ambiguous facial expressions. According to Ekman (1994, p. 15), the 'use of the term basic is to emphasize the role that evolution has played in shaping both the unique and the common features that emotion display, as well as its current function', but it is questionable that emotions should be restricted to only those with prominent display (facial) features: 'display' should be expanded to include behavioural functions (e.g., risk assessment).

which is unavoidable (e.g., realizing you have a terminal illness). He likened this blending to that observed in colour perception: we perceive a vast number of colours from only three types of cones in the retina that are maximally sensitive to electro-magnetic energy of a given frequency; and the wonderful variety of colours seen on television is achieved with only three types of colour pixels. At this point, it should be noted that *how* the brain achieves this blending of basic emotional states to form complex ones is not known; far from being a limitation specific to RST, this problem represents one of the fundamental 'mysteries' in brain science.

However, some 'basic' emotions may not be as important as the name implies. For example, disgust is not likely to be a major emotional and motivational factor of general influence, as it is restricted to avoiding contaminated and rotting food – although by associative learning it can be linked to conditioned stimuli (e.g., one religion finding another religion's food preference as 'disgusting'; cf. Pinker 1997). In this respect, it is perhaps useful to distinguish between general systems of emotion and motivation (incentive and avoidance systems; FFFS and BAS) and those systems dealing with environmental demands in the form of response-specific processes (e.g., disgust and nausea).

There is also the real problem of essentially the same emotions being labelled with different terms, depending on the specifics of the situation. For example, McDonald and Leary (2005) make a strong case for considering the emotion associated with social exclusion as comparable to that associated with pain. Corr (2005) addresses this issue from the standpoint of RST and argues that, under some conditions (e.g., extreme psychological strain), differently labelled emotions (e.g., associated with distress and pain) may well be highly similar (if not identical), although this homology may break down under different conditions (e.g., mild psychological strain). An important challenge for future RST research will be to show just how far the blending theory of FFFS, BIS and BAS goes in understanding the multiple emotions that exist (see Table 1.2). Nonetheless, as pointed out by Matthews (chapter 17), there is a need to relate RST processes to specifically social brain processes (e.g., FFFS/BIS and attachment styles).

In the case of depression, a common example is the death of a loved one. The 'stimulus' – here the complex stimuli in memory and the environment relating to this person – cannot be avoided (i.e., forgotten) and thus is approached (via recurring thoughts, conversations with relatives, etc.). In this case, the stimulus (i.e., the person) elicits not the emotion of anxiety but an emotion that is usually called 'sadness'. As we shall see in chapter 2, therapeutic drugs effective for anxiety are also

Table 1.2 *Emotions/states and behaviours associated with: (a) the avoidance of (FFFS) and approach (BIS) to aversive stimuli, and (b) the approach to appetitive stimuli*

	Stimulus Conditions	Emotion/State	Behaviours
Aversive stimuli			
Avoid (FFFS):	Avoidable	Fear	Phobic avoidance, Escape, Flight
	Unavoidable	Panic	Fight (defensive aggression), Freeze
Approach (BIS):	Avoidable	Anxiety	Behavioural inhibition, Risk assessment
	Unavoidable	Depression	Behavioural suppression
Appetitive stimuli			
Approach (BA)	Attainable	Hope, Anticipatory pleasure	Exploration, Sub-goal scaffolding
	Unattainable	Frustration, Anger	Fight (predatory aggression), Displacement activity

effective for depression, pointing to a close (but not homologous) association between these two states that seem prima facie unrelated. When this state entails thoughts about one's past behaviour, then sadness may be tinged with the emotion of regret.

This line of reasoning may help to throw light upon some very peculiar human states and beliefs. For example, in the above case of the death of a loved one, imagine the effect of this dead person, somehow, miraculously coming back to life: this would be an extreme form of 'relief of non-punishment', which is an input to the BAS. Given the power of the human brain to generate advanced states of fantasy, we may speculate that it is this very emotion that drives (i.e., provides the primary positive reinforcement in Skinnerian terminology) the hope that, one day, we shall all be reunited with dead loved ones – this is a very common belief, codified in many religions both primitive and advanced. Seen in the light of RST, we may start to understand why sadness is often accompanied by BAS-related 'hope' – for this reason, the emotions associated with bereavement tend to be complex. As this example shows, the blending of even only three major systems of emotion can give rise to many and complex emotional states, and by extension can be applied to explicating even fundamentally human beliefs.

As we can also see from Table 1.2, some appetitive stimuli that are being approached are unattainable – this fact might not be evident at the outset of the BAS sub-goal scaffolding process. This outcome would be related to a state of 'frustrative non-reward', which itself is an adequate input to the FFFS. Depending on its intensity it may produce fight/aggression and anger – the theoretical rationale for this process is given by Corr (2002b). Other authors have also argued that the BAS is related to anger (Carver 2004; Harmon–Jones 2003). Perhaps depending on personality factors outside the BAS (e.g., the *Violence Inhibition Mechanism*; see Blair, Jones, Clark and Smith 1995) we might also expect to see predatory aggression under conditions that signal reward unattainment, as well as various forms of displacement activity.

Cognition

One persistent criticism of RST is that it fails to consider adequately the importance of cognitive processes in emotion and personality (see Matthews, chapter 17). Revised RST has gone some way to remedying this situation. According to Gray and McNaughton (2000), the septo-hippocampal system is involved in cognitive and memorial processing, and this theory predicts that pathological anxiety is likely to result, at least in some cases, from abnormal cognitive processing. This brings revised RST much closer to recent cognitive theories of anxiety (e.g., Mathews 1993; Eysenck 1992). However, there are deeper criticisms of the form of the 'biological' approach adopted by RST, some of which may represent a misunderstanding, or a failure on behalf of RST researchers to be sufficiently clear in their theorizing. RST does not purport to offer a biological theory of personality that, in some way, supersedes or circumvents the need for the consideration of cognitive constructs; and nor does RST see cognitive processes as epiphenomenal froth – indeed, as argued below, it is possible to construct a theory that reveals the explicit role of higher-level cognitive constructs. For the moment, it may be noted that the influence of 'knowledge' is important – what, after all else, is a conditioned stimulus if not a form of knowledge about the relationship between stimuli? What RST does emphasize is that, whatever represents the eliciting stimuli for fear and anxiety – and these, of course, are influenced by primary and secondary appraisal and knowledge-level representations (see below) – the immediate behavioural reactions *must* be mediated by neural systems specifically evolved to control involuntary and fast-action processes. Irrespective of what we might consider the primary influence on

anxiety – be it a specific threat stimulus, or the words on this page (that must require high-level cognitive processing and the engagement of knowledge structures) and which may be unique to the individual concerned (e.g., intense fear elicited by the sight of pink blancmange) – RST assumes that your *experience* of fear and anxiety is the same as everyone else's: these emotions are *not* themselves knowledge-level representations of symbolic interactions. According to this position, the type of constructs considered by RST, as well as the wider family of biological-level constructs in personality, are indispensable to a full account of emotion and personality: they are *necessary* processes. The extent to which RST possess *sufficient* processes to provide a full account of emotion and personality is a somewhat different issue.

We can see the problem more clearly in the following way. Any account that supposes that fear and anxiety result exclusively from information processing, unrelated to low-level neural systems in the brain, would have a difficult (impossible?) task of providing a cogent theoretical account, especially one that would explain the remarkable commonality observed in behaviour between non-human and human animals (e.g., effects of anxiolytics; see chapters 2 and 5). In contrast to some cognitive approaches, RST poses the question: why should we assert that, for example, the high fear induced in the rat when confronted suddenly and unexpectedly by a cat and the human being when attacked by a vicious predator – including the typical act of defecation (common in soldiers in combat) – is so qualitatively different so as to demand entirely different explanatory constructs, even when this emotion is affected by the same drugs? RST *does* need to invoke different explanatory constructs when talking about the factors *sufficient* to activate these primitive neural modules: the details of the knowledge structures and cognitive processes that activate *your* emotions do not, of necessity, need to be the same as those that activate *mine*, yet our emotional experience will be highly similar (in as far as we can ever make the general statement about inter-individual similarities, and between-species similarities, the former of which show commonality in verbal report, behaviour and reactions to drugs, as do the latter, save verbal report, although even non-human animals vocalize as part of the fear response repertoire).

None of the above is designed to detract from the important role played by cognitive factors (e.g., expectancy, primary and secondary appraisal, and 'knowledge' level factors): these play a central role in determining the adequate inputs to the FFFS, BAS and BIS; accordingly, they serve to regulate emotions and defensive behaviour: if inputs are changed then outputs of these systems are also changed.

Coping

Coping is one area where the forging of RST and purely cognitive approaches could start. Hasking (2006) considered this question in the context of eating and drinking behaviours, noting that both reinforcement sensitivity and coping strategies have independently been related to these behaviours. Hasking hypothesized that if sensitivity to reward and punishment are biological predispositions that regulate behaviour, then it may be assumed that reinforcement sensitivity is a *distal* predictor of behaviour. On the other hand, coping strategies may be seen as the *proximal* predictor of behaviour. Hasking goes on to argue that these distal and proximal factors may be related in at least two ways. (a) High BIS sensitivity leads a person to adopt more avoidant coping strategies, such as denial, in order to avoid negative stimuli associated with the stressor, and the use of avoidant coping may in turn predict dysfunctional eating patterns or alcohol use. (b) Coping strategies may influence the relationship between reinforcement sensitivity and eating or drinking behaviour: there may be a positive relationship between BAS sensitivity and alcohol use, but only for people who engage in avoidant coping strategies, while a negative relationship may exist for those who use problem-focused strategies. In such a way, the coping strategies a person chooses may either mediate or moderate the relationship between predispositional reinforcement sensitivity and eating or drinking behaviour (and, by inference, many other forms of behaviour). The potential importance of a distinction between *distal* and *proximal* levels of explanation has been noted by other RST researchers (e.g., Jackson and Francis 2004).

Between-species and inter-individual differences

The question of the relevance of non-human animal data for human personality is often asked (see McNaughton and Corr, chapter 3); and the conclusion is sometimes drawn that any theory that relies so heavily upon animal data must have limitations when applied to complex human psychological processes. But RST assumes that neural systems of emotion and motivation are not species-specific; they are shared by a large variety of species. However, the specific demands of each species are quite different. RST assumes that the neural systems provide the general, background, evolutionary foundations to avoid (FFFS), approach (BAS) and be cautious (BIS), but the *specific stimuli* that we avoid, approach and are cautious of have important species-specific features.

The (relatively) strong BAS activation and (relatively weak) FFFS activation of a conference presenter should produce a cautious, risk assessment approach to preparation (e.g., checking for spelling mistakes in handout). A chimpanzee does not show these behaviours. He is not concerned with conference presentations; his concerns lie elsewhere (e.g., 'presenting' his genetic fitness to mates). But his concerns are no less reliant on basic emotional and motivational systems; and so too are ours. The fact that one species, in this case human beings, seem to have basic needs for competence, autonomy, social connection, even 'spiritual transcendence', is irrelevant to assessing the validity of RST, which does not purport to explain every 'basic need' of every species. Rather, it attempts to provide explanatory constructs that work at the general level, referring to general influences – irrespective of the specific content of the needs – related to avoidance, approach and cautious of, whatever set of stimuli dominate at the species level or individual level. Thus, the specific details of the life challenges facing a species, or individuals within a species, are not of primary importance. They, of course, need to be taken into account when trying to work out the types of stimuli which are sufficient to activate basic systems in a given species or individual.

Experimental assays

A number of putative experimental assays of the FFFS, BAS and BIS are shown in chapter 5 (Table 5.1). A challenge for RST is to develop 'pure' measures of these systems; i.e., measures that allow the threshold and activation of the three systems to be measured without the influence of other systems. Only once such measures have been developed will it be possible to ask and answer questions about the interplay of these systems. In order to develop such paradigmatic assays, it will be necessary to undertake a detailed task analysis to define the parameters of the task and how these are likely to be affected by RST processes. In this respect, it would be useful to move beyond a verbal-qualitative description to a numerical-quantitative one using computational modelling procedures that capture the dynamics of behaviour and allow experimental predictions to be generated using different parameter values. (Examples of this approach are given in chapters 5 and 16.) Such computer simulations often produce results counter to verbal-qualitative description, and accordingly may help to explain the divergence of results reported in the RST literature (Corr 2004).

Much of previous RST research has not taken adequate account of the cognitive and behavioural demands of the task. In the 2000 revised theory, we can see that associated with each defensive distance is a

specific set of cognitions/behaviours (e.g., flee vs. freeze). Even when there are significant RST-related individual differences on the task, the pattern of effects may confuse this relationship because of the match/mismatch between task demands and specific RST cognitions/behaviours. A related problem concerns the range of effects: manipulation of reinforcement at only one point on the range (which is common in RST research) may be affecting performance on only one 'limb' of the performance curve. If this curve is inverted-U then sensitivity to reinforcement may be *either* positively related or negatively related to performance. This possibility may explain the diversity of results in the RST literature.

Levels of control

Discussion of the role of cognitive factors in RST processes raises a general problem of the appropriate level of control in emotion and motivation: 'biological' or 'cognitive'? Presenting the problem in this binary form is not helpful. Instead, what is needed is a model that clarifies the roles played by each level of control. The aim of this section is to embed RST in a wider literature which has already considered many of the issues of behavioural control. In so doing, the relevance – indeed, crucial importance – of considering consciousness in personality research will be highlighted.

Consciousness: what is controlling behaviour?

It is perhaps surprisingly that the nature of consciousness is all too rarely discussed alongside emotion and personality and, hitherto, never in relation to RST. However, this is not unique to personality research, as the problems of consciousness, especially those that seem so scientifically intractable, have, at least until the recent past, been largely ignored (or, more often, unrecognized).

We are, therefore, fortunate that Gray's (2004) last book, *Consciousness: Creeping up on the Hard Problem*, addressed the problems of consciousness for psychology in general, especially the problem of the relationship between systems controlling behaviour and conscious awareness. It is likely that, as with many other areas opened up by Gray's thinking, this hitherto delinquent area of psychology will come increasingly within the spotlight of personality psychology. (Space prevents a thorough discussion of this topic; for a more detailed description of Gray's model, see Corr 2006.) First, Gray does not offer an account of the 'Hard Problem' (Chalmers 1995), i.e., the *why* and *how* of conscious experience, especially

how the brain *generates* conscious awareness. It instead addresses the *function* of consciousness: what it is for and how is it implemented?

'Online' and 'Offline' processes In addressing the function of consciousness, the distinction between different levels of behavioural control is important. Standard psychology textbooks continue to contrast 'learning theories' and 'cognitive theories'; and this approach follows the long-fought territorial battles between stimulus-response (S-R) theories (e.g., Skinner), who argued for automatic bonds between eliciting stimuli and responses, and cognitive theorists (e.g., Tolman), who argued that intervening variables between stimuli and responses, knowledge structures and processes are required (see MacPhail 1998). In reviews of the literature, Toates (1998, 2006) draws attention to the fact that both processes are observed in human and non-human animals, and that consideration of both processes may help us better to understand normal and abnormal behaviour in general, and consciousness in particular. This debate is also played out in the literatures concerned with implicit/procedural and explicit/declarative processes. In the context of personality research, some of the major dual-process theories of behavioural control are well rehearsed by Carver (2005).

A similar distinction is seen in the field of visual perception. Milner and Goodale (1995) refer to two streams in visual processing: the 'online' and 'seeing' systems. The online system, is the 'action system' which can be indexed by various performance measures; it is automatic and reflex-like, occurring before the time needed to achieve conscious awareness of the action and the eliciting stimuli. This system seems to use the dorsal processing stream – Milner and Goodale propose that rather than being the 'where' stream, as suggested by Ungerleider and Mishkin (1982), it is the 'how' stream. The ventral stream, in contrast, is largely conscious: it is the 'what' stream, and is similar to the 'offline' processes discussed in this section.

According to Toates' (1998) model (Figure 1.3), a stimulus (*S*) has a certain strength of tendency to produce a response (*R*; formerly called 'habit strength'); i.e., *S* has a response eliciting potential, which varies from zero to some maximum value (this strength depends upon innate factors and learning). 'Cognition' in this context refers to those processes that encode knowledge about the world in a form not tied to particular behaviours (*R*s). Where there is uncertainly, novelty or a mismatch of actual against expected outcomes, behavioural control shifts from the S-R (online) processing, which is fast and coarse-grained in its analysis, to cognitive (offline) processing, that is slow and

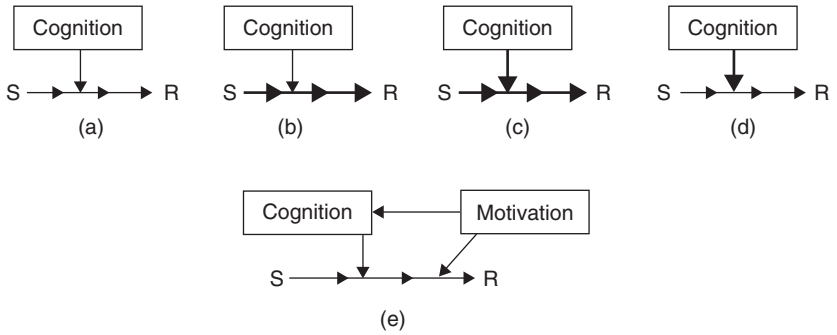

Figure 1.3 Representation of relationship between S–R and cognitive processes: (a) balanced weighting; (b) strong S–R and weak cognition; (c) strong S–R and strong cognition; (d) weak S–R and strong cognition; and (e) interaction with motivation

fine-grained in its analysis. The particular circumstance that gives rise to the different weightings is shown in Figure 1.3.

Toates' model helps us to understand the adaptive value of consciousness. This model contends that some actions that can be organized at a low (online) level can nonetheless be affected by conscious (offline) processes: higher-order, conscious processes can modulate the strength of connections controlling behaviour organized at a relatively low level of the control hierarchy. For example, a fear state that is processed consciously may sensitize the whole defensive system and thereby affect *subsequent* fast, automatic responses (such as the startle reflex). Thus, Toates' emphasizes the weights attached to motor programs, and how cognitive, conscious (offline) processes can modify the weights (i.e., firing potential) of online responses. Online processes correspond to an open feedback system; whereas offline processes correspond to a closed feedback system (see above).

Gray's (2004) theory of the function of consciousness extends the line of argument pursued by Toates. Specifically, Gray takes seriously the implications of the findings of Libet (1985, 2003) which have shown a number of rather counterintuitive phenomena. Various forms of data point to the fact that it takes some 300–500 ms of brain activity for consciousness to occur: this is the 'lateness' of conscious experience.

The problem with such findings for any adaptive theory of consciousness is that long before 300–500 ms, motor actions have already been initiated (e.g., the removal of the hand from a hot stove occurs before awareness of the hand touching the stove). In this specific case, removal of the hand is involuntary and not controlled by conscious

processes. However, events are not experienced as if they happened 300–500 ms ago: consciousness appears to refer to what is happening *now*. Libet suggests that the conscious experience of a stimulus is 'referred back in time' once neuronal adequacy has been achieved to make it *seem* as if there was no delay. This produces the illusion of voluntary control; arguably, it is an illusion that continues to dominate views on the role of cognition in personality.[10]

Now, there have been many criticisms of Libet's experiments as well as his interpretation of his data (e.g., Libet 2003; Zhu 2003; see Blackmore 2003), but the basic finding of the lateness of conscious awareness seem solid. As noted by Gray (2004, p. 23), 'The scandal of Libet's findings is that they show *the conscious awareness of volition to be illusory*' (emphasis added). However, from a physicalist point of view, all mind events (e.g., thinking and consciousness) must be *caused* by a physical process in the brain that *precedes* the conscious awareness of these events – how could it be otherwise? But, this leaves us with the problems of causation and behavioural control.

What does all of this mean for RST? Well, it implies that conscious awareness of emotion, volition, behaviour, etc. does not play any role in the emotion, volition and behaviour *to which it refers*. Now, we must be careful not to conflate 'cognitive processes' and 'conscious awareness'; but it remains the case that what we are consciously aware of does not have an *immediate* causal role to play – but we shall shortly see it does exert causal effects on *subsequent* behaviour. The main point is that the volition of behaviour is *always* non-conscious in terms of its *direct*, or primary, causal process. Thus, according to this position, RST relates to immediate causal processes, leaving higher-level cognitive processes (e.g., verbal mediation, but not exclusively so) to *indirect*, or secondary, causal effects on future (if only hundreds of milliseconds) behaviours.

Function of conscious awareness: late error detection According to Gray (2004, p. 107):

> Conscious experience serves three linked functions. (1) It contains a model of the relatively enduring features of the external world; and the model is experienced as though it *is* the external world; (2) within the framework afforded by this model, features that are particularly relevant to ongoing motor programs or which depart from expectation are monitored and emphasised; (3) within the

[10] Space prevents a full exposition of this matter. It is sufficient to say that 'cognition' does not relate solely to conscious processes, or necessarily only to slow, fine-grained analysis; but it is also germane to point out that fast-reflexive actions cannot entail much in the way of complex process *at the time* of execution of the response.

framework of the model, the controlled variables and set-points of the brain's unconscious servomechanisms can be juxtaposed, combined and modified; in this way, error can be corrected.

To understand these functions, imagine you are confronted by a dangerous snake and your fear system fires off an automatic (online) motor program: all this happens long before (i.e., hundreds of milliseconds) you are consciously aware of (i.e., 'see' and 'feel') the snake. It would now be highly adaptive to 'replay' the immediate past in order to analyse its contents, especially at those times when the online fear behaviour did not achieve its goal (in this instance, increasing defensive distance).

Central to this model of conscious awareness is the 'comparator',[11] which, in RST, serves to compare actual stimuli with expected stimuli (Gray and Smith 1969). Thus, the comparator compares the *expected* state of the world with the *actual* state of the world. When there is no discrepancy, and 'all is going to plan', the comparator is said to be in 'just checking mode'; however, when there is a mismatch between the expected and actual states of the world, then the comparator goes into 'control mode' (Gray 1981). According to Gray, in this control mode, the *contents* of consciousness are generated (e.g., attention to snake).

The relevance of online and offline systems can now be seen. According to this model, online (non-conscious) processes are modified by offline (conscious) processes; in Toates' terminology, the weights attached to response propensities in online processes are adjusted on the basis of the fine-grained offline processes. Gray (2004) uses the terminology of cybernetics with behavioural weights attached to specific stimuli (see Figure 1.4).[12]

Now, offline processes do have causal effects on *subsequent* online processes; in other words, our behaviour is modified by experience: we *learn*. Before our discussion slides blindly into a dualistic mode of thinking, it needs to be emphasized that both online and offline processes are products of the brain, but they have different functions. Specifically, they differ in (a) their temporal characteristics; (b) their level of analysis; and (c) their representation in conscious awareness. Online processes are

[11] Carver and Scheier (1998) put forward the intriguing idea that focusing attention on the self is often equivalent to engaging the comparator, which is centrally involved in self-regulating feedback control processes in general. In Gray's model, the comparator compares expected and actual reinforcing stimuli; in the case of personality, this leads to rumination, worry and anxiety; but in the case of consciousness, when behaviour is not going to plan (i.e., a feedback error signal is generated) offline processes are triggered (*ex hypothesi*, consciousness).

[12] Cybernetics is the science of communication and control, comprising end-goals and feedback processes containing control of values within the system that guide the organism towards its final goal (Wiener 1948).

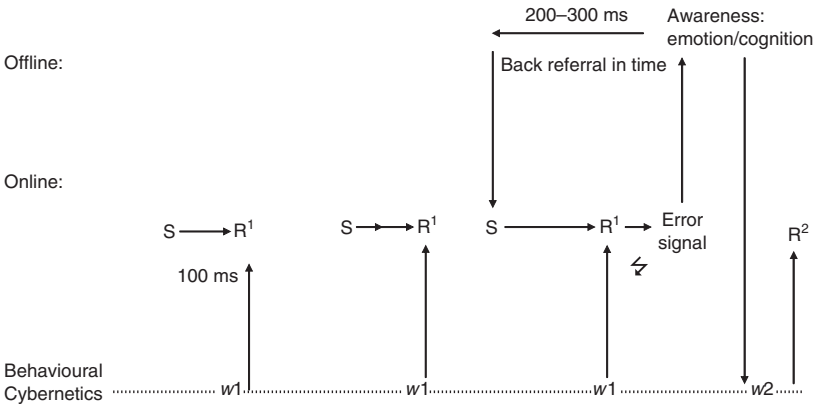

Figure 1.4 A schematic depiction of Gray's late error detection model of consciousness. Offline processes monitor the success of (automatic) online processes, and when 'everything is going to plan', online processes 'fire-off' and are not influenced by offline processes; but when an error signal (ϟ) is detected, then the salient features of the error signal are transferred to offline processes which, among other things, is represented in the form of conscious perceptual experience (both perceptual in terms of imagery, etc. and affective, in terms of emotion). Although offline processes, of necessity, lag behind on-line processes, the process of 'back referral in term' provides the illusion that the experience is occurring at the same moment as the stimuli that it represents. Of importance, this offline process cannot affect responses to the stimuli it represents, but it can alter the behavioural weights of *subsequent* online processes (e.g., speed of response) and thus exerts a causal effect on future behaviour ('once bitten, twice shy')

very fast, involving coarse-grained analysis of salient features, and are not represented in conscious awareness; in contrast, offline processes are slow, taking hundreds of milliseconds to generate, entail (relatively) fine-grained analysis, and are represented in conscious awareness. This dual system serves evolutionary challenges well: a fast 'dirty' response system coupled with a slower 'cleaner' response system for post-action processing. (The distinction between 'fast-coarse', not involving conscious processing, and 'fine-detailed', involving conscious awareness, levels of processing has been noted before; LeDoux 1996, 2000.)

Thus, online behaviour, which *always* comes before the generation of conscious awareness, can be modified by offline processing that brings to the fore those salient features (e.g., novelty and mismatch) that require closer analysis. Although we may still want to talk about 'fast-and-dirty' primary cognitive analysis of data, we may also talk about

'cybernetic behavioural weights' that are central to RST – certainly, on this analysis, we cannot talk about conscious primary analysis, while secondary appraisal, that may be conscious, is not causally efficacious at all in relation to the stimuli it is appraising.

In relation to the BIS, one important consequence of modifying behavioural weights attached to online processes is to inhibit prepotent (online) responses. This mechanism solves one major evolutionary problem: how to ensure that online automatic responses are appropriate. It would be desirable to be able to inhibit the firing-off of these automatic behavioural routines in some circumstances (e.g., inhibiting avoidance behaviours when in foraging mode), even if this inhibition takes several hundreds of milliseconds (usually enough time to have important consequences). Gray notes that conscious control is exerted only at critical junctures, when a definite choice has to be made. But when in automatic mode, errors can occur. Thus, automatic routines are well suited to fixed tasks – crucially for RST, fixed action patterns associated with defensive distance and environmental parameters – but they are not so good for tasks requiring a departure from fixed routines (e.g., a novel task) or when automatic performance is not going to plan.

Extending the late detection model: what-if simulations

Consistent with the general form of Gray's (2004) model is the additional idea that consciousness allows 'what-if' simulations of future behaviour, produced offline in a virtual reality environment that contains the important features of the real physical environment (e.g., imagination). Indeed, this function seems highly important to human beings: much of our time is spent *imagining* the likely consequences of our behaviour and making plans for the future. Such behaviours require complex computational processes, specifically involving inferences concerning the likely behaviour of other people. It is obvious that personality does relate to such simulations, and, it may be speculated, that much of the 'energy' for neurotic disorders comes from these 'offline' cognitive processes. It is thus likely that variance in neuroticism will not be solely explained by standard RST processes, but by an elaboration of the theory to incorporate longer-term cognitive processes that decouple these processes from actual stimuli: thus, we can have 'free-floating' anxiety and 'worry about worry'. In reality it is highly probable that these higher-level effects are, themselves shaped by basic RST processes in early development and are, in any event, refuelled by ongoing FFFS, BIS ad BAS activation (or deactivation in the case of deficit syndromes, e.g., introvertive-anhedonia).

Let us now put to work our offline, what-if simulation modeller. Compare two phenotypes facing complex social problems: one phenotype simply computes errors in online programs; the second anticipates these errors by running realistic simulations, away from the (potentially punishing) stimuli in the real environment (informed by previous outcomes of online behaviours). Which phenotype would be more successful in terms of survival and reproduction? In many environments, it would be the latter. Now, if this same process could be used to solve complex social problems associated with approach motivation (e.g., securing and keeping a mate, influencing other conspecifics, achieving status and privileges in the social structure) then it is possible to conceive of the development of behaviour that is appropriate to complex natural and social environments that require a combination of fast, reflex-like processes, as well as more reflective analysis of the environment. Thus, the offline, late error detection mechanism seems to have acquired the capacity to be dissociated from immediately preceding online processes. This additional evolution of offline processing may well necessitate awareness (i.e., the experience of qualia, e.g., the 'redness' of a rose), especially when adequate sensory stimuli are not present in order to build an *apparently* real-time simulation model encompassing the central features of the external world.

It is tempting to associate FFFS, BIS and BAS with rapid 'online' behavioural procedures, and more 'frontal' systems, associated with restraint and deliberative, attentional control, with cognitive procedures of reflection and modification. If this approach has any merit then we might be able to understand how factors such as psychoticism[13] (in Eysenck's model) and constraint (in Tellegen's model) appear to be independent of neuroticism, anxiety and approach behaviour, yet seem able to modify these more basic reactions. However, it should be borne in mind that the above account of behavioural control provides only a very rough sketch; further elaboration and refinement of these processes is needed. In particular, it will be necessary to account for cognitive processes that operate automatically as it will be for those that operate at a more controlled level entailing slow deliberation and fine-grained analysis.

[13] Relating psychoticism to a higher level of cognitive control – that, as we have seen, according to Gray's theory is responsible for generating the contents of consciousness – is consistent with the association of psychoticism with the cognitive processes disrupted in schizophrenia (see Gray *et al.* 1991); of course, schizophrenia itself represents a disruption of conscious experience of self and the world. According to the theory promulgated here, the disruption seen in schizophrenia is, in part, a failure of this higher-level system to provide adequate constructions and inferences of the internal and external stimuli which are *re*presented, or displayed, in the form of conscious awareness.

Conclusion

RST will ultimately be tested against the criterion of *progressive science* (Lakatos 1970): will it only be interested in parochial theoretical issues, fortified by a restricted range of 'special case' data, or does it hold the potential to throw light on existing theoretical vistas as well as opening up new ones in personality psychology? Forging links with other theoretical perspectives will be especially crucial, as well as expanding the spheres of interest (e.g., occupational and financial behaviour). In the final analysis, it must be borne in mind that all scientific theories are, at best, approximations to the natural phenomena they attempt to explain; and if they are not to become conceptual fossils, they must change and grow.

References

Abler, B., Walter, H., Erk S., Kammerer, H. and Spitzer, M. (2006), Prediction error as a linear function of reward probability is coded in human nucleus accumbens, *NeuroImage* xx 31, 790–795

Avila, C. (2001), Distinguishing BIS–mediated and BAS-mediated disinhibition mechanisms: a comparison of disinhibition models of Gray (1981, 1987) and of Patterson and Newman (1993), *Journal of Personality and Social Psychology*, 80, 311–324

Barlow, D.H. (1988), *Anxiety and its Disorders* (New York: Guilford Press)

Blanchard, R.J. and Blanchard, D.C. (1990), An ethoexperimental analysis of defense, fear and anxiety in N. McNaughton and G. Andrews (eds), *Anxiety* (Dunedin: Otago University Press), pp. 12–133

Blackmore, S. (2003), *Consciousness: An Introduction* (London: Hodder and Stoughton)

Blair, R.J.R., Jones, L., Clark, F. and Smith, M. (1995), Is the psychopath 'morally insane'? *Personality and Individual Differences*, 19, 741–752

Carver, C.S. (2004), Negative affects deriving from the behavioral approach system, *Emotion*, 41, 3–22

(2005), Impulse and constraint: perspectives from personality psychology, convergence with theory in other areas, and potential for integration, *Personality and Social Psychology Review*, 9, 312–333

Carver, C.S. and Scheier, M.F. (1998), *On the Self-Regulation of Behavior* (Cambridge: Cambridge University Press)

Carver, C.S. and White, T.L. (1994), Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales, *Journal of Personality and Social Psychology*, 67, 319–333

Chalmers, D.J. (1995), Facing up to the problem of consciousness, *Journal of Consciousness Studies*, 2, 200–219

Cloninger, C.R. (1986), A unified biosocial theory of personality and its role in the development of anxiety states, *Psychiatric Developments*, 3, 167–226

Corr, P.J. (2001), Testing problems in J.A. Gray's personality theory: a commentary on Matthews and Gilliland (1999), *Personal Individual Differences*, 30, 333–352

Corr, P.J. (2002a), J.A. Gray's reinforcement sensitivity theory: tests of the joint subsystem hypothesis of anxiety and impulsivity, *Personality and Individual Differences*, 33, 511–532

(2002b), J.A. Gray's reinforcement sensitivity theory and frustrative nonreward: a theoretical note on expectancies in reactions to rewarding stimuli, *Personality and Individual Differences*, 32, 1247–1253

(2004), Reinforcement sensitivity theory and personality, *Neuroscience and Biobehavioral Reviews*, 28, 317–332

(2005), Social exclusion and the hierarchical defense system: comment on MacDonald and Leary (2005), *Psychological Bulletin*, 131, 231–236

(2006), *Understanding Biological Psychology* (Oxford: Blackwell)

Corr, P.J. and Perkins, A.M (2006), The role of theory in the psychophysiology of personality: from Ivan Pavlov to Jeffrey Gray, *International Journal of Psychophysiology*, 62, 367–376

Corr, P.J., Pickering, A.D. and Gray, J.A. (1995a), Personality and reinforcement in associative and instrumental learning, *Personal Individual Differences*, 19, 47–71

(1995b), Sociability/impulsivity and caffeine-induced arousal effects: critical flicker/fusion frequency and procedural learning, *Personality and Individual Differences*, 18, 713–730

Darwin, C. (1859/1968), *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (Princeton, NJ: Princeton University Press)

Dawkins, R. and Krebs, J.R. (1979), Arms races between and within species, *Proceeding of the Royal Society of London*, Series B, 205, 489–511

Dickman, S.J. (1990), Functional and dysfunctional impulsivity: personality and cognitive correlates, *Journal of Personality and Social Psychology*, 58, 95–102

Eilam, D. (2005), Die hard: a blend of freezing and fleeing as a dynamic defense: implications for the control of defensive behaviour, *Neuroscience and Biobehavioral Reviews*, 1181–1191

Ekman P. (1994), All emotions are basic in P. Ekman and R.J. Davidson (eds), *The Nature of Emotion: Fundamental Questions* (Oxford: Oxford University Press) pp. 15–19

Eysenck, H.J. (1957), *The Dynamics of Anxiety and Hysteria* (New York: Preger)

(1967), *The Biological Basis of Personality* (IL: Thomas, Springfield)

(1979), The conditioning model of neurosis, *Behavioural and Brain Sciences*, 2, 155–199

Eysenck, H.J. and Eysenck, M.W. (1985), *Personality and Individual Differences: A Natural Science Approach* (New York: Plenum Press)

Eysenck, H.J. and Levey, A. (1972), Conditioning, introversion–extraversion and the strength of the nervous system in V.D. Nebylitsyn and J.A. Gray (eds), *Biological Bases of Individual Behaviour* (London: Academic Press, London), pp. 206–220

Eysenck, M.W. (1992), *Anxiety: the Cognitive Perspective* (Hillsdale, NJ: Erlbaum)

Eysenck, S.B.G., Pearson, P.R., Easting, G. and Allsopp, J.F. (1985), Age norms for impulsiveness, venturesomeness and empathy in adults, *Personality and Individual Differences*, 6, 613–619

Fowles, D.C. (2000), Electrodermal hypoactivity and antisocial behaviour: does anxiety mediate the relationship, *Journal of Affective Disorders*, 61, 177–189

Fowles, D.C. (2006), Jeffrey Gray's contributions to theories of anxiety, personality, and psychopathology in T. Canli (ed.), *Biology of Personality and Individual Differences* (New York: Guilford Press), pp. 7–34

Gray, J.A. (1970), The psychophysiological basis of introversion–extraversion, *Behaviour Research and Therapy*, 8, 249–266

(1972a), Learning theory, the conceptual nervous system and personality in V.D. Nebylitsyn and J.A. Gray (eds), *The Biological Bases of Individual Behaviour* (New York: Academic Press), pp. 372–399

(1972b), The psychophysiological nature of introversion-extraversion: a modification of Eysenck's theory in V.D. Nebylitsyn and J.A. Gray (eds), *The Biological Bases of Individual Behaviour* (New York: Academic Press), pp. 182–205

(1975), *Elements of a Two-Process Theory of Learning* (London: Academic Press)

(1976), The behavioural inhibition system: a possible substrate for anxiety in M.P. Feldman and A.M. Broadhurst (eds), *Theoretical and Experimental Bases of Behaviour Modification* (London: Wiley), pp. 3–41

(1981), A critique of Eysenck's theory of personality in H.J. Eysenck (ed.), *A Model for Personality* (Berlin: Springer), pp. 246–276

(1982), *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System* (Oxford: Oxford University Press)

(1985), Anxiety and the brain: pigments aren't colour names, *Bulletin of the British Psychological Society*, 38, 299–300

(1987), *The Psychology of Fear and Stress* (Cambridge: Cambridge University Press)

(1994), Three fundamental emotion systems. In P. Ekman and R.J. Davidson (eds), *The Nature of Emotion: Fundamental Questions* (Oxford: Oxford University Press)

(2004), *Consciousness: Creeping up on the Hard Problem* (Oxford: Oxford University Press)

Gray, J.A., Feldon, J., Rawlins, J.N.P., Hemsley, D.R. and Smith, A.D. (1991), The neuropsychology of schizophrenia, *Behavioral and Brain Sciences*, 14, 1–84

Gray, J.A. and N. McNaughton (2000), *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System* (Oxford: Oxford University Press)

Gray J.A. and Smith, P.T. (1969), An arousal decision model for partial reinforcement and discrimination learning in R.M. Gilbert and N.S. Sutherland (eds), *Animal Discrimination Learning* (London: Academic Press), pp. 243–272

Gomez, R., Cooper, A. and Gomez, A. (2005), An item response theory analysis of the Carver and White (1994) BIS/BAS Scales, *Personality and Individual Differences*, 39, 1093–1103

Hasking, P.A. (2006), Reinforcement sensitivity, coping, disordered eating and drinking behaviour in adolescents, *Personality and Individual Differences*, 40, 677–688

Harmon-Jones, E. (2003), Anger and the behavioural approach system, *Personality and Individual Differences*, 35, 995–1005

Hebb, D.O. (1955), Drives and the CNS (conceptual nervous system), *Psychological Review,* 62, 243–254

Hull, C.L. (1952), *A Behavior System* (New Haven: Yale University Press)

Jackson, C.J. and Francis, L.J. (2004), Are interactions in Gray's reinforcement sensitivity theory proximal or distal in the prediction of religiosity: a test of the joint subsystems hypothesis, *Personality and Individual Differences*, 36, 1197–1209

Jackson, C.J. and Smillie, L.D. (2004), Appetitive motivation predicts the majority of personality and an ability measure: a comparison of BAS measures and a re-evaluation of the importance of RST, *Personality and Individual Differences*, 36, 1627–1636

Konorski, J. (1967), *Integrative Activity of the Brain* (Chicago: Chicago University Press)

Lakatos, I. (1970), Falsification and the methodology of scientific research programmes in I. Lakatos and A. Musgrave (eds), *Criticism and the Growth of Knowledge* (Cambridge: Cambridge University Press), pp. 91–196

LeDoux, J.E. (1996), *The Emotional Brain* (New York: Simon and Schuster)

(2000), Emotion circuits in the brain, *Annual Review of Neuroscience*, 23, 155–184

Libet, B. (1985), Unconscious cerebral initiative and the role of conscious will in voluntary action, *Behavioral and Brain Sciences*, 8, 529–566

(2003), Timing of conscious experience: Reply to the 2002 commentaries on Libet's findings, *Consciousness and Cognition*, 12, 321–331

MacDonald, G. and Leary, M.R. (2005), Why does social exclusion hurt? The relationship between social and physical pain, *Psychological Bulletin*, 131, 202–223

Mackintosh, N.J. (1983), *Conditioning and Associative Learning*. (Oxford: Clarendon Press)

MacPhail, E.M. (1998), *The Evolution of Consciousness* (Oxford: Oxford University Press)

Mathews, A. (1993), Biases in processing emotional information, *The Psychologist*, 6, 493–499

McNaughton, N. and Corr, P.J. (2004), A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance, *Neuroscience and Biobehavioral Reviews*, 28, 285–305

Miller. G.A., Gelenter. E. and Pribram, K. (1960), *Plans and the Structure of Behavior* (New York: Holt, Rinehart and Winston)

Milner, A.D. and Goodale, M.A. (1995), *The Visual Brain in Action* (Oxford: Oxford University Press)

Mowrer, H.O. (1960), *Learning Theory and Behavior* (New York: Wiley)

Olds, J. and Milner, P. (1954), Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain, *Journal of Comparative and Physiological Psychology*, 47, 419–427

Patterson, C.M. and Newman, J.P. (1993), Reflectivity and learning from aversive events: towards a psychological mechanism for the syndromes of disinhibition, *Psychological Review*, 100, 716–736

Pavlov, I.P. (1927), *Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford: Oxford University Press, G.V. Anrep (trans and ed.))

Perkins, A.M. and Corr, P.J. (2006), Reactions to threat and personality: psychometric differentiation of intensity and direction dimensions of human defensive behaviour, *Behavioural Brain Research*, 169, 21–28

Phillips, M.L., Williams, L.M., Heining, M., Herba, C.M., Russell, T., Andrew, C., Bullmore, E.T., Brammer, M.J., Williams, S.C.R. and Morgan, M.J. (2004), Differential neural responses to overt and covert presentations of facial expressions of fear and disgust, *Neuroimage*, 21, 1484–1496

Pickering, A.D., Corr, P.J., Gray, J.A. (1999), Interactions and reinforcement sensitivity theory: a theoretical analysis of Rusting and Larsen (1997), *Personality and Individual Differences*, 26, 357–365

Pinker, S. (1997), *How the Mind Works* (New York: Norton)

Ranter, S.C. (1977), Immobility in invertebrates: what can we learn? *Psychological Review*, 1, 1–14

Toates, F. (1998), The interaction of cognitive and stimulus-response processes in the control of behaviour, *Neuroscience and Biobehavioral Reviews*, 22, 59–83

(2006), A model of the hierarchy of behaviour, cognition, and consciousness, *Consciousness and Cognition*, 15, 75–118

Ungerleider, L.G. and Mishkin, M. (1982), Two cortical vision systems in D.J. Ingle, M.A. Goodale and R.J.W. Mansfield (eds), *Analysis of Visual Behaviour* (Cambridge, MA: MIT Press)

Weiner, N. (1948), *Cybernetics, or Control and Communication in the Animal and the Machine* (New York: John Wiley & Sons)

Zhu, J. (2003), Reclaiming volition: an alternative interpretation of Libet's experiment, *Journal of Consciousness Studies*, 10, 61–77