21 The Reinforcement Sensitivity Theory of Personality

Philip J. Corr

> 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.
>
> (Jeremy Bentham, *Introduction to the Principles of Morals and Legislation* (1781))

In one form or another, Bentham's 'masters' of pain and pleasure remain the sovereign of behaviour, and underpin the moral and judicial framework of all societies. We have yet to document a society where behaviour is governed by the dominant pursuit of pain and the avoidance of pleasure – for sure, there are organizations (e.g., the Roman Catholic *Opus Dei*) where mortification, entailing physical pain, is sanctioned (indeed, in this example, sanctified); but, typically, these relatively mild forms of suffering are in the service of a greater pleasure (e.g., eternity in Heaven). Moving from the spiritual to the temporal plane, day-to-day life is regulated by striving for the good things (e.g., safety, food, drink and fulfilling social, personal and occupational pursuits), as well as the avoidance of bad things (e.g., dangerous animals, rotting food and criticism from other people) – that is, 'goods' and 'bads' in the nomenclature of rational economics. In our personal life, the power and ubiquity of these 'sovereign masters' is such that we rarely have the need to reflect upon them: they are accepted 'givens' of everyday life, even though they populate much of our conscious awareness, and in psychopathological conditions (e.g., Obsessional-Compulsive Disorder) dominate it. Their importance was recognized by twentieth century academic psychology, which was dominated by Behaviourism, with its focus on the role of reinforcement (positive and negative) and punishment in shaping behaviour (and the mind more generally), as well as the early philosophers of Ancient Greece (e.g., Epicures of Samos 341–270 BC, and Aristotle 384–322 BC). In other realms of life, such as the penal-justice system, considerations of 'right' and 'wrong' often reduce to questions of how best to design behavioural control instruments that, it is hoped, deter transgression of legal codes.

We may, therefore, sensibly enquire after a scientific theory that helps us to understand the psychology of the control of behaviour based on these sovereign masters; and we may also wonder why these sovereign masters are so often implicated in aberrations of normally-regulated behaviour, expressed in the variety of forms of psychopathology (e.g., the affective disorders and various addictions). Moreover, we may wonder as to the evolutionary foundations of these regulatory forces, and how they give rise to individual differences in the underlying neuropsychological systems that comprise 'personality' (Corr 2007). Indeed, we may go further to enquire as to the role they play in consciousness, where these sovereign masters are often found to exert their influence. This chapter discusses these issues in the context of one major neuropsychological theory that attempts to account for the influence of pain and pleasure on the variety of factors that compose human behavioural choreography.

Foundations of Reinforcement Sensitivity Theory

The Reinforcement Sensitivity Theory (RST)[1] of personality represents a bold attempt to account for the neuropsychological regulation of behaviour, and how individual differences in neuropsychological systems give rise to what we commonly label 'personality'. RST is based upon notions of central states of emotion and motivation that mediate the relations between stimulus input and behavioural response: here 'stimulus' and 'response' can be internal processes, and only inferred from ingenious behavioural experiments (e.g., sensory preconditioning; see McNaughton and Corr 2008).

In this section, I summarize the development of Jeffrey Gray's (1970, 1975, 1976, 1982) neuropsychological theory of emotion, motivation, learning and personality, which is now widely known as RST. Although it will be seen that much of the analysis of behaviour follows standard procedures used in behavioural psychology, as well as many of the experimental tools of the behaviourist, the explanatory framework is very different to that of the strict behaviourist, most famously B. F. Skinner who considered central states of emotion, etc. as wrong-headed causal 'fictions' (Skinner 1953). Stimuli per se do not affect behaviour (at least, in any simple sense); they merely have the potential to activate neuropsychological systems (i.e., internal processes) that control behavioural reactions: the mind is not a series of black boxes.[2] For a fully-satisfying scientific

[1] As noted by one of the originators of the name, 'Reinforcement Sensitivity Theory' (Pickering, Diaz and Gray 1995), Pickering (2008) considered alternative names: 'Reinforcement Reactivity Theory' and 'Motivational Input Sensitivity Theory'. Reinforcement 'sensitivity' is arguably the best choice as it does not require that activation of the systems will always be evident in overt, and directly measurable, behavioural reactions.

[2] The power of behavioural techniques, when stripped of related explanatory framework, provide the best behavioural evidence for the existence of central states of emotion and motivation; we see examples of this in the case of 'frustrative non-reward' and 'relief of non-punishment', which are, in strict behaviourist terms, non-events. Their effects only make sense if we infer central states of expectation, suggesting an internal comparator that compares expected and actual motivationally-significant inputs. For example, frustrative non-reward effects are seen in the partial reinforcement

explanation of behaviour control and regulation, it is to these neuropsychological systems that we must turn our attention.

RST evolved over the past forty years, from its inception in 1970, and it has gone through several refinements, most notably by Gray and McNaughton (2000). As we shall see throughout this chapter, RST can appear, at first blush, complex, indeed confusing, because it encompasses a number of approaches that move at different paces. This point is well made by Smillie, Pickering and Jackson (2006, p. 320), who note that, although RST is often seen as a theory of personality, it is 'more accurately identified as a neuropsychology of emotion, motivation and learning. In fact, RST was born of basic animal learning research, initially not at all concerned with personality'. The fact that RST is an evolving theory is a strength (i.e., it is 'progressive' theory; see Lakatos 1970); however, this state of flux makes it something of a moving target for personality researchers, 'as if it were frozen in time, Gray's "personality model" is a relatively discrete slice of an otherwise continuous and ongoing field of knowledge' (Smillie, Pickering and Jackson 2006, p. 321). As we shall see below, this problem can be much reduced by separating RST into its *state* and *traits* components.

Another important aspect of RST is the distinction between those parts that belong to the *conceptual nervous system* (cns) and those parts that belong to the *central nervous system* (CNS) (a distinction advanced by Hebb 1955). The cns component of RST provides the behavioural scaffolding, formalized within some theoretical framework (e.g., learning theory; see Gray 1975; or, ethoexperimental analysis; see Gray and McNaughton 2000); the CNS component specifies the brain systems involved, couched in terms of the latest knowledge of the neuroendocrine system (see McNaughton and Corr 2008). As noted by Gray (1972a), these two levels of explanation *must* be compatible. Thus, we can talk of a *neuropsychology* of behaviour, as well as the effects of individual differences in the operating parameters of these systems that give rise to 'personality'. Gray used the language of cybernetics (cf. Weiner 1948) – 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 – in the form of a cns-CNS bridge, to show how the flow of information and control of outputs is achieved (see also, Gray 2004).

Identification and clarification of emotion and motivation systems

Before delving into the details of RST, it is important to appreciate the logic that underlies Gray's approach to science. In common with other theorists, Gray faced two major problems: first, how to identify brain systems responsible for behaviour; and, secondly, how to characterize these systems once identified.

extinction effect (PREE), the effects of which do not find cogent explanation in terms of non-emotional learning (see Fowles 2006).

The individual differences perspective is one major way of identifying major sources of variation in behaviour; by inference, there must be causal systems (i.e., sources) giving rise to observed variations in behaviour. Hans Eysenck's (1947, 1957, 1967) approach was to use multivariate statistical analysis to identify these major sources of variation in the form of personality dimensions. Gray accepted that this 'top-down' approach can identify the minimum *number* of sources of variation (i.e., the 'extraction problem' in factor analysis), but he argued that such statistical approaches can never resolve the correct *orientation* of these observed dimensions (i.e., the 'rotation problem' in factor analysis). Gray's alternative 'bottom-up' approach to identifying major systems of causal influence rested on other forms of evidence, including the effects of brain lesions, experimental brain research (e.g., intracranial self-stimulation studies), and, of most importance, the effects on behaviour of classes of drugs known to be effective in the treatment of psychiatric disorders: this was Gray's 'philosopher's stone' – transforming base pharmacological findings into a valuable neuropsychological theory. This was a subtle and clever way to expose the nature of fundamental emotion and motivation systems, especially those implicated in major forms of psychopathology.

Gray argued the following: if we want to know what is the brain-behavioural nature of 'anxiety' (the scary quotes here reflect the fact that the phenomenon to be explained has received only a partial and rather superficial description), then we can pursue the following course of action: (a) take drugs that are effective against human anxiety (i.e., those psychological disorders recognized as falling under the rubric of 'anxiety'); then (b) analyse their behavioural profile in non-human animals to understand their more fundamental nature; and then (c) compare these behavioural profiles with other drugs (e.g., psychostimulates). Thus, by a careful analysis of the behavioural effects of different *classes* of drugs (e.g., anxiety vs. psychostimulates), a detailed description may be formed of the underlying systems – the assumption that these different behavioural effects reflect different underlying systems follows standard neuroscientific reasoning (see Corr 2006).

Gray reasoned that anxiolytic (i.e., anti-anxiety) drugs provided a criterion for what constitutes anxiety. Gray (1977) provided an exhaustive review of the behavioural effects of minor tranquilizers (i.e., barbiturates, alcohol and benzodiazepines, which at that time were the dominant class of anxiolytic drugs) on the following behavioural paradigms: rewarded behaviour, passive avoidance, classical conditioning of fear, escape behaviour, one-way active avoidance, two-way active avoidance, responses elicited by aversive stimuli, and frustrative non-reward (as seen in resistance to extinction), discrimination learning, intermittent reinforcement schedules in the Skinner box, reduction of reward and the after-effects of reward. The reasoning proceeds that once a behavioural dissection has been achieved, based on behavioural reactions to classes of drugs, then it is much easier to identify actual neuropsychological systems that these drugs act upon. Following the emphasis of behavioural psychology on overt behaviour, Gray did not favour a research strategy based on a purely human and verbal source of information (e.g., self-reports of patients), but one that could be tested, via rigorous experimentation, in non-human animals: the goal of identifying the neural substrate for anxiety was, and largely still is, only possible with the use of experimental animals. Gray's whole theoretical approach rests and falls on these major assumptions.

The major findings from Gray's (1977) exhaustive review of the behavioural effects of anxiolytic drugs were: they anatagonize or reduce the behavioural effects (i.e., suppression of behaviour) associated with conditioned stimuli for punishment (Pun-CSs) and frustative non-reward (nonRew-CSs; i.e., the non-delivery of expected reward), as well as, but less strongly, novel stimuli. Noteworthy, was the relative absence of effects on behaviour controlled by *unconditioned* punishing or rewarding stimuli (i.e., innate stimuli). As discussed below, this evidence suggested that anxiolytic drugs acted on a system that was responsible for *behavioural inhibition* in reaction to conditioned signals of punishment, non-reward (frustration) and novelty.

States and traits

RST is built upon a description of the immediate/short-term *state* of neural systems: how animals, including the human form, respond to motivationally significant (i.e., 'reinforcing') stimuli, and which neuropsychological systems mediate these responses. Built upon this state infrastructure are longer-term *trait* dispositions of emotion, motivation and behaviour. As we move to psychopathology, we see the role played by both factors. Figure 21.1 shows a conceptual framework that illustrates these different processes.

RST assumes that personality factors revealed by multivariate statistical analysis (e.g., factor analysis) reflect sources of variation in neuropsychological systems that are stable over time – that is, they are properties of the individual. Personality traits account for behavioural *differences* between individuals presented with identical environments, and, also, the consistency of behaviour seen in any one individual over time. According to this position, the ultimate goal of personality research is to identify the relatively stable biological (i.e., genes and neuroendocrine systems) variables that determine the factor structure that is 'recovered' from statistical analysis of behaviour (including verbal output and checking boxes on personality questionnaires; Corr 2004; Corr and McNaughton 2008; McNaughton and Corr 2004). This theoretical position is not to deny the importance of the environment in controlling behaviour (for example, the environment seems to determine whether depression or anxiety is expressed in individuals with the same genes for internalizing disorders; e.g., Kendler, Prescott, Myers and Neale 2003; see below). However, to produce consistent long-term effects, environmental influences must be instantiated in biological systems: environmental influences do not have any substance unless there is a biological system to mediate them.

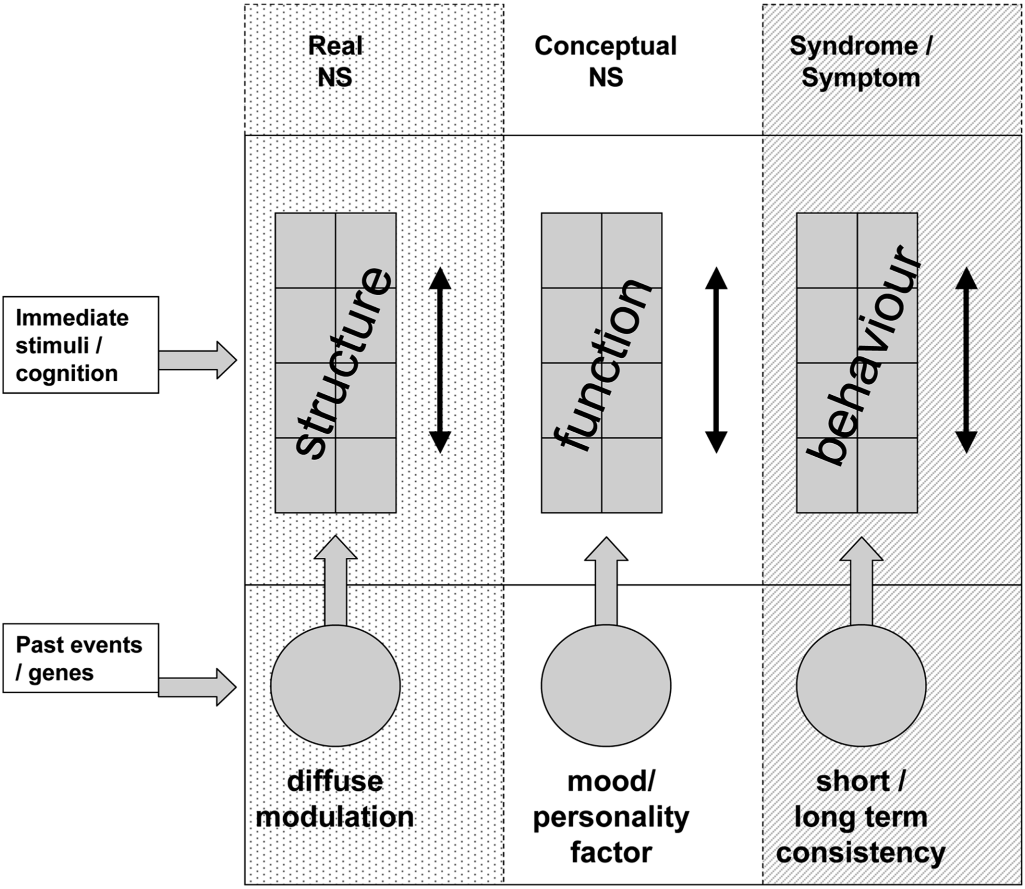

Figure 21.1. *The relationship between (a) the real nervous system (Real NS), (b) the conceptual nervous system (Conceptual NS), (c) syndromes/behaviours related to (d) immediate stimuli/cognitions, and (e) past events/genes, providing descriptions in terms of structure, function and behaviour.*

Development of Reinforcement Sensitivity Theory

RST has gone through several phases of development. In the sections to follow, I concentrate on the theory as it exists in 2009. However, a brief 'Cook's Tour' of the milestones in RST's development is necessary in order to appreciate how the current theory developed (for a fuller discussion, see Corr 2008a).

The 'necessity' handmaiden to the mother of the invention of RST was the need to resolve the gross cracks that appeared in the major biological personality theory of that time, namely, Hans Eysenck's (1967) arousal/activation theory of Introversion-Extraversion (E) and Neuroticism (N). Eysenck's 'top-down' approach consisted in first 'discovering' the major dimensions of personality, and, secondly, providing a theoretical (biological) account for their existence. But, as discussed elsewhere (Corr and McNaughton 2008), multivariate statistical analysis is unable to 'recover' the separate causal influences that get conflated in immediate/short-term behaviour responses, as well as in the longer-term development of personality: what is measured in behaviour is the net products of, possibly separate, causal influences and the operation of their underlying systems. What Eysenck seemed to have found were major *descriptive* dimensions of personality (principally, E and N), that reflect the causal influences of separate, and interacting, underlying systems, and which, as such, could only ever be tied to very general biological processes that cut across these underlying systems, specifically neuropsychological arousal and activation, of the ascending reticular activating system (ARAS) and visceral system, respectively (for a summary, see Corr 2004). Given the fundamental limitation of multivariate statistical techniques of extraction (e.g., factor analysis; see Lykken 1971), arguably Eysenck's approach never stood a decent chance of unravelling the complexity of underlying biological systems. This fact alone may well account for the multiple cracks that rapidly appeared in his theoretical edifice (see Gray 1981). As Gray's RST is usually seen as a development and refinement of Eysenck's *general* approach, we now need to turn to the specific details of Eysenck's theory to see the problems that Gray attempted to solve.

Hans Eysenck's personality theory

Eysenck's (1967) personality theory states that individuals differ with respect to the sensitivity of their ARAS, which serves to dampen or amplify incoming sensory stimulation. Those of us with an active ARAS easily generate cortical arousal, whereas those of us with a less active ARAS generate cortical arousal much more slowly. It was assumed (but no theoretical rationale was given for this) that there exists an optimal level of arousal: too little or too much leads to poor hedonic tone, which motivates us to alter this sub-optimal arousal state. According to this view, those of us with an overactive ARAS are, generally, more cortically aroused and closer to our optimal point of arousal; therefore, we do not seek out more stimulation, and we shy away from stimulation that we encounter: we are introverts. In contrast, those of us with an underactive ARAS are, generally, less cortically aroused and are not close to this optimal point of arousal; therefore, we seek out more stimulation, and we benefit from stimulation that we encounter: we are extraverts. Most people are in the middle range of these extreme values (i.e., ambiverts). What we measure in personality questionnaires are these preferences and behaviours.

Inspired by Pavlov's theory of excitatory and inhibitory brain processes being associated with conditioning (a theory capitalized upon in Eysenck's 1957 theory), Eysenck stated that introverted individuals (i.e., high arousal, or excitable process, type) are relatively easy to condition; whereas, extraverts (i.e., low arousal, or inhibitory process, type) are relatively less easy to condition. The observation that clinical neurotics are indeed introverts (they are also high on neurosis, which adds negative emotional fuel to the high-arousal fire) fitted the theory well, as did the clinical observation that behaviour therapy, which was based upon conditioning principles, was effective in the treatment of a number of neurotic conditions. Such was the elegance and wide-range explanatory power of Eysenck's theory, it became highly influential and widely accepted – it was seen as a *tour de force* in personality-psychopathology research. Alas, ugly data – including Eysenck's very own – was to ruin this beautiful theory.

The first problem was that, at high levels of stimulation, introverts were actually *worse* than extraverts at conditioning (Eysenck and Levey 1972). Although this supported the Pavlovian notion of transmarginal inhibition (TMI) of response (i.e., a breakdown of the orderly stimuli-response relationship at too-high levels of stimulation), it simultaneously corroded the very foundations of the theory, for it led to the conclusion that extraverts should condition best to high arousing stimuli (including the panoply of aversive stimuli found in neurosis) and, therefore, should be overrepresented in the psychiatric clinic, which they are not for typical neurotic conditions.

Secondly, compounded with this first problem was the finding, again from Eysenck's own work (Eysenck and Levey 1972) but also from other researchers (Revelle 1997), that it is impulsivity, not sociability, that carried the causal burden of the arousal-conditioning link. As impulsivity is orthogonal, and thus independent of sociability (the main trait of Eysenck's Extraversion scale), this destroyed not only the arousal-conditioning-Extraversion link, but also the relevance of Extraversion at all in conditioning effects, including those supposedly so crucial in the development of neurotic conditions.

If these two problems were not enough to destroy finally Eysenck's already tarnished theory, thirdly, the relations observed between arousal and conditioning were observed to vary as a function of time of day: Eysenck-like sociability/impulsivity x arousal effects that are found with morning testing (e.g., introverts showing superior performance under placebo and TMI-related performance deficits under arousal, relative to extraverts) are reversed with evening testing. As ruefully noted by Gray (1981), one is not a neurotic in the morning and a psychopath in the evening!

While these findings pointed to the power of general arousal theory, at the same moment they undermined the particulars of Eysenck's personality theory.[3]

However, worse still was to follow. *Even if we assume* that Eysenck's theory were correct, classical conditioning cannot account for the known phenomena

[3] It is not too fanciful to propose the following in defence of Eysenck's theory. First, most aversive conditioning of children is during the earlier part of the day (i.e., during school hours); secondly, much aversive stimulation is relatively mild; and thirdly, and perhaps of most importance, conditioning entails an incubation period (Eysenck 1979) consisting of rehearsal in memory of the aversive experience, over extensive periods of time, during states of lower arousal. As shown below, the Extraversion-arousal link may still be a viable part of personality theory, including RST (e.g., how initially neutral stimuli get conditioned in the first place).

of neurosis. As discussed by Corr (2008a), the classical conditioning theory of neurosis assumes that, as a result of the conditioned stimulus (CS) (e.g., hairy animal) and unconditioned stimulus (UCS) (e.g., pain of dog bite) getting paired, the CS comes to take on the eliciting properties of the UCS, such that, after conditioning and when presented alone, the CS produces a response (i.e., the conditioned response (CR), e.g., fear, and its associated behaviours) that resembles the unconditioned response (UCR) (e.g., pain, and its associated behaviours) elicited by the UCS. All well and good thus far (assuming that 'fear' is equivalent to 'pain', which itself is something of a leap of faith). But, there is a major problem with this theory. The CR (e.g., fear) *does not* substitute for the UCR (e.g., pain). In some crucial 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 – recall the last time an object hit you hard!); but these reactions are quite different – in fact, opposite to – a CS *signalling* pain, which consists of a different range of behaviours (e.g., quietness and behavioural inhibition). A lingering problem here concerns emotion: where does fear come from? More technically, where is fear generated in the brain, and how is this fear-system related to conditioning? Eysenck seemed just to assume that emotion arose spontaneously; but this simply will not do. In addition, if there is a fear generating system, then maybe that is where we should look for the genesis of clinical neurosis.

Another clue to the potential importance of an innate fear system was the debate between Eysenck's and Spence's laboratories where, in the latter, it was found that conditioning was related to anxiety not (low) Extraversion. This debate was finally resolved by the realization that it is anxiety related to conditioning in laboratories that is more threatening (as in the case of Spence's; Spence 1964). This realization was accepted by Eysenck as a satisfactory resolution to this empirical difference; however, it could have occurred to him, as it did to Gray later, that the very resolution was bought at the cost of an even greater problem: what led to the greater threat-related conditioning in Spence's laboratory? Emotion was never satisfactorily explained in Eysenck's theory: it was seen, at varying times, as a cause (e.g., in Spence's conditioning studies), as an outcome (e.g., in neurosis), and as a regulatory set point mechanism (e.g., in arousal and hedonic tone relations). In Eysenck's theory, it remained something of an unruly, even delinquent, construct.

Jeffrey Gray's reward and punishment systems

As a former doctoral student of Eysenck's, and much later as the successor to his Departmental Chair at the Institute of Psychiatry in London, Gray was well aware of his former mentor's theory, as well as the deep roots it had in Pavlovian psychology and in the relatively newer Hullian learning theory and neurophysiology (e.g., Gray 1964). This knowledge allowed Gray not just to criticize Eysenck's personality theory, but to dismantle its theoretical foundations, especially the focus on one system of drive/arousal that was fundamentally Hullian in

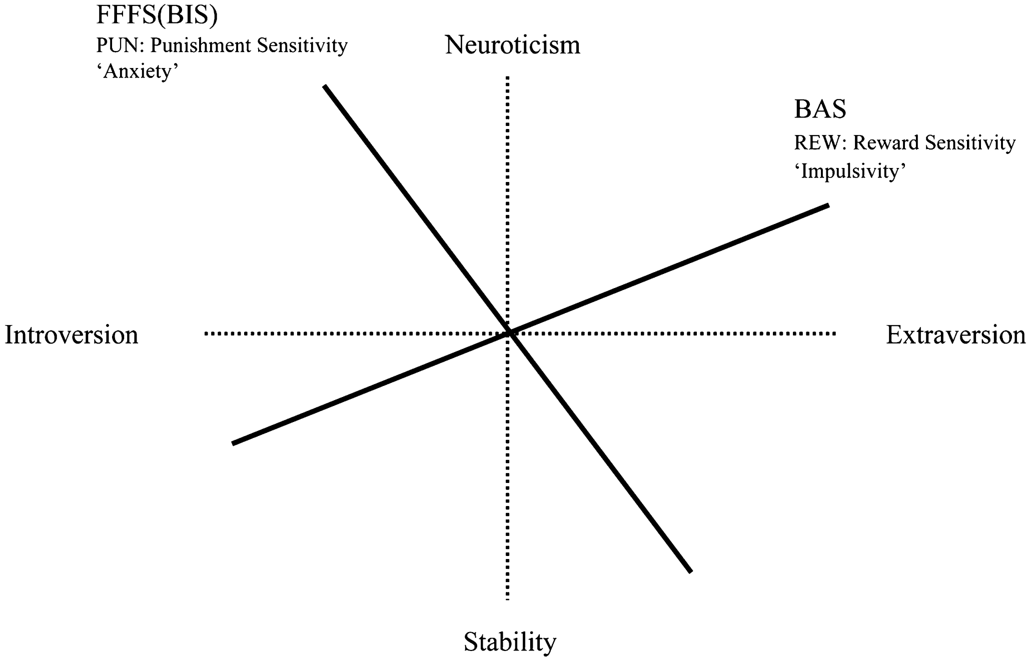

Figure 21.2. *Position in factor space of the fundamental punishment sensitivity and reward sensitivity (unbroken lines) and the emergent surface expressions of these sensitivities, i.e., Extraversion (E) and Neuroticism (N) (broken lines). In the revised theory (see text), a clear distinction exists 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.*

nature (see Corr 2008a; Corr, Pickering and Gray 1995). In its place came a two-process theory of learning, entailing separate dimensions of reward and punishment, a focus on the fundamental role of internal states of emotion, and a much more sophisticated neuropsychology.

In brief, Gray (1970, 1972b, 1981) proposed a modification of Eysenck's theory thus: (a) to the position of Extraversion (E) and Neuroticism (N) in multivariate statistical factor space; and (b) to their neuropsychological bases. According to Gray, E and N should be rotated, approximately, 30° to form the more causally efficient axes of 'punishment sensitivity', reflecting Anxiety (Anx), and 'reward sensitivity', reflecting Impulsivity (Imp) (Figure 21.2).

Gray's modification stated that highly impulsive individuals (Imp+) are most sensitive to *signals* of reward, relative to their low impulsive (Imp−) counterparts;[4] and highly anxious individuals (Anx+) are most sensitive to *signals* of punishment, relative to low anxiety (Anx−) counterparts. It was assumed that Imp

[4] The notion that impulsivity, which has its high pole in the neurotic-extravert quadrant of E/N space, was related to reward came from several sources of evidence: (a) two-factor learning theory

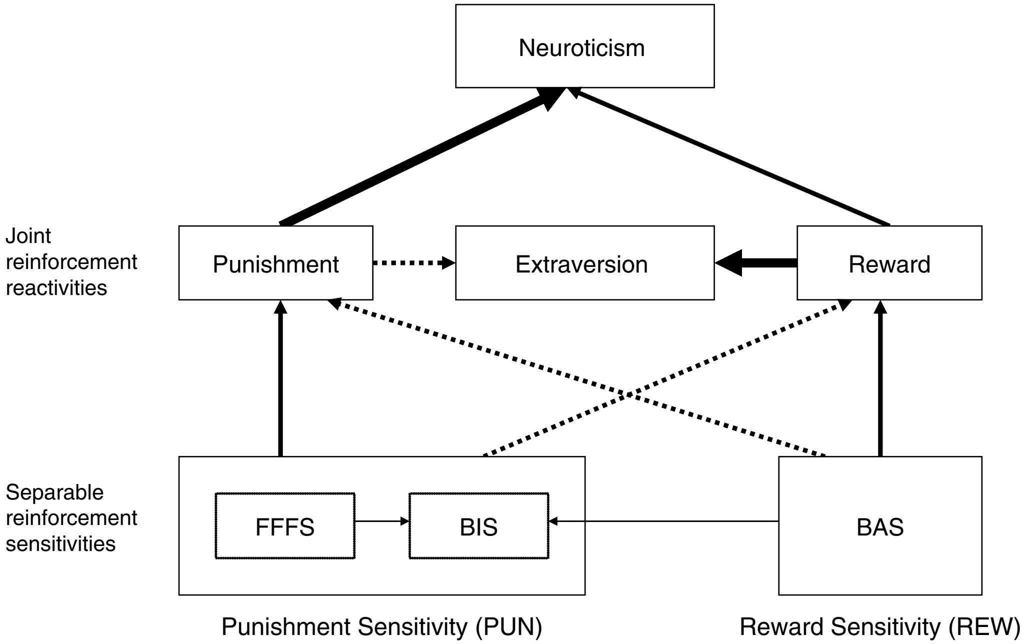
Neuroticism
Joint reinforcement reactivities
Punishment
Extraversion
Reward
Separable reinforcement sensitivities
FFFS
BIS
BAS
Punishment Sensitivity (PUN)
Reward Sensitivity (REW)

Figure 21.3. *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 (see Figure 21.2): 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.*

and Anx, and their processes, were independent – this position is now known as 'separable subsystems hypothesis' (Corr 2001, 2002a; see Corr and McNaughton 2008). According to this new view, Eysenck's E and N dimensions are secondary (conflated) factors of these more fundamental traits/processes (see Figure 21.3).

(Mowrer 1960; Konorski 1967), that showed that behavioural reactions to aversive stimuli are controlled by a different system to that controlling behavioural reactions to appetitive stimuli; (b) the *relative* insensitivity of anxiolytic drugs to affect behavioural reactions to appetitive stimuli; and (c) the psychological data showing that highly impulsive people are more prone to engage in a variety of 'sociopsychiatric' behaviours (e.g., gambling, and other 'externalizing disorders' of an extraverted and sociable nature).

Gray's (1970) theory deftly side-stepped the problems accompanying Eysenck's, and it also explained *why* introverts were, generally, more cortically aroused: they are more punishment sensitive (punishment is more arousing than reward); and, as extraverts are more sensitive to reward, not punishment, they are, accordingly, less aroused. In addition, Gray (1970) argued that drugs that reduce clinical anxiety lower N and raise E scores, as does psychosurgery to the frontal cortex (whether caused by accident or surgical design) – both sets of findings suggest that a single anxiety dimension is a better account than two, separate, dimensions.

Two factor learning theory

Lurking behind these theoretical developments were advances being made in learning theory. As already noted, Eysenck's theory followed in the tradition of Hullian (1952) learning theory, which reduced all forms of motivationally-salient reinforcement to a single process of 'drive-reduction'; 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'. However, at this time, there was a strong movement away from Hull's grand theory of behaviour – which has now fallen by the wayside of science – towards a two factor theory of learning based upon reward and punishment systems. It was Mowrer's (1960) seminal work that contributed to this development: he argued that the effects of reward and punishment had different behavioural effects, as well as different underlying bases, and he specifically introduced the notion that central states of emotion (e.g., 'hope') mediate stimuli and responses. For a mediation to occur, there must be a mediating system. These general ideas entered mainstream psychology through the writings of such people as Konorski (1967) and Mackintosh (1983). Gray's (1975) *Elements of a two-process theory of learning* fully embodied this tradition in personality psychology.[5] On the real nervous system side of the coin, the conceptual nervous system work was strengthened by neurophysiological findings pointing to specific emotion centres in the brain (e.g., the 'pleasure centres'; Olds and Milner 1954; see Corr 2006).

From these converging lines of evidence, Gray (1970) advanced the claim that the 'emotions' are elicited by motivationally-significant ('reinforcing') stimuli (of any kind) that activate innate systems in the brain. Now seen as rather innocuous, this claim has important and widespread implications for personality psychology: if emotion, and its related motivation, were fundamental to personality (as suggested by Eysenck's own work in linking personality to psychopathology) then we may better understand personality by understanding emotion systems in the brain.

In critiquing Eysenck's approach, Gray noted that classical conditioning does not, indeed cannot, create emotion, normal or pathological; all it can do is to

[5] For a rebuttal of the claim (widely held, if not so frequently articulated) that non-human behaviour/cognition is irrelevant to our understanding of human emotion, motivation and personality, see McNaughton and Corr (2008b).

transform initially neutral stimuli into conditioned (reinforcing) stimuli that, via Pavlovian classical conditioning, acquire the power to activate innate systems of emotion which, themselves, are responsible for generating emotion. Thus, according to this position, reduction of pathological emotions can be achieved in one of two ways: (a) deconditioning aversive reinforcing stimuli, which weakens the strength of stimulus inputs into the innate emotion systems; or (b) by dampening down the activity in the systems themselves (e.g., by the use of drugs that target key molecules in parts of the innate system). We may see the effectiveness of cognitive-behavioural therapy (CBT) as another way to ‘decondition’ the power of hitherto aversive stimuli to activate the emotion systems (e.g., by restructuring ‘irrational’ cognitions that serve as inputs into these systems).

Two broad affective dimensions

We have now covered the main conceptual and developmental parts of the evolution of Gray’s RST, which we can summarize in the words of Fowles (2006, p. 8):

> In this view, organisms are seen as maximizing exposure to rewarding (‘appetitive’) events and minimizing exposure to punishing (‘aversive’) events. Rewarding or appetitive events consist of the presentation of a reward (Rew), termination of a punishment (Pun!), or omission of an expected punishment (nonPun), while punishing or aversive events consist of the punishment (Pun), termination of reward (Rew!), and omission of an expected reward (nonRew). Through a process of classical conditioning, conditioned stimuli (CSs) paired with events come to acquire some of their emotional and motivational properties.

An important point to note here is the fact that reward (Rew) itself and the termination of a punishment (Pun!) or omission of an expected punishment (nonPun; relief of non-punishment), share much in common in terms of their functions and pharmacology; and in a complementary way, punishment (Pun) itself and the termination of reward (Rew!), and omission of an expected reward (nonRew; ‘frustrative non-reward’), are similarly common. Somewhat unique to RST, this analysis draws attention not to observed behaviour but to the internal, central states that underlie them. It is at this deeper level of analysis that we see the operation of core psychological processes (McNaughton and Corr 2008).

Summary of Pre-2000 RST

We now know that the anxiety system was characterized on the basis of a detailed analysis of the pattern of behavioural effects of classes of drugs known to affect anxiety in human beings (mainly barbiturates and the benzodiazepines (Gray 1977), later to be extended to novel anxiety reducing drugs, i.e., novel anxiolytics; see below). This detailed analysis (summarized in Gray 1982) led to the formal definition of the BIS.

(1) The behavioural inhibition system (BIS) was postulated to be sensitive to *conditioned* aversive stimuli, omission/termination of expected reward, and conditioned frustration (i.e., conditioning to stimuli that signalled expected reward, non-reward), as well as an assortment of other inputs, including extreme novelty, high intensity stimuli and innate fear stimuli (e.g., snakes, blood). This system was charged with suppressing ongoing operant behaviour in the face of threat, which allowed for enhanced information-processing and vigilance. The BIS was related to the personality factor of Anxiety (Anx). The neural instantiation of the BIS was postulated to be in the septo-hippocampal system of the brain.

According to Gray, anxiolytic drugs work by impairing the activity of the BIS and thus its outputs, making behaviour less risk averse and, colloquially speaking, less concerned (worried) with potential sources of danger. Although anxiety was associated with BIS activity, its phenomenological nature was not considered, and it is still unclear how and where this subjective state is generated (this problem is not restricted to Gray's theory, but to all subjective experiences; see below).

(2) The fight-flight system (FFS) was postulated 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 speculatively associated by Gray with Eysenck's personality factor of Psychoticism (P) (Eysenck and Eysenck 1976). The neural instantiation of the FFS was postulated to be in the periaqueductal grey and (various nuclei of) the hypothalamus.

(3) The behavioural approach system (BAS) was postulated 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 state positive affect (PA) and the personality dimension of Impulsivity (Imp). The neural instantiation of the BAS was postulated to be in the mesolimbic dopamine circuit.

The experimental evidence testing the pre-2000 theory was summarized by a review paper, (Corr 2004) and an edited book (Corr 2008b) that surveyed all the main areas of RST.

Post-2000 RST

Gray and NcNaughton (2000) substantially revised BIS theory and RST more generally. This revision updates and elaborates the older theory and, crucially in some important respects, makes different predictions (for more detailed discussion of these matters, see Corr 2004, 2008a; Corr and McNaughton 2008; McNaughton and Corr 2004, 2008a).

Revised RST, once again, postulates three systems.

(1) The fight–flight–freeze system (FFFS) is now responsible for mediating reactions to *all* aversive stimuli, conditioned and unconditioned. It updates the FFS to include 'freezing' (see below). In addition, the theory proposes a

hierarchical array of neural modules, each responsible for a specific defensive behaviour (e.g., avoidance and freezing). The FFFS mediates the emotion of fear, not anxiety. The associated personality factor consists of fear-proneness and avoidance, which clinically may be mapped onto such disorders as phobia and panic. This is the 'Get me out of here!' system.

(2) The behavioural approach system (BAS) mediates reactions to *all* appetitive stimuli, conditioned and unconditioned, and is the least changed of the three systems. It interfaces with dedicated consummatory systems (e.g., eating and drinking) which are responsible for the final consumption of unconditioned stimuli (e.g., food); the BAS is involved in the incentive processes moving the animals up the temporo-spatial gradient to the final biological reinforcer. It is responsible for generating the emotion of 'anticipatory pleasure', and hope itself. The associated personality factor consists of optimism, reward-orientation and (especially in very high BAS-active individuals) impulsiveness (but see below), which clinically may be mapped onto addictive behaviours (e.g., pathological gambling) and various varieties of high-risk, impulsive behaviour. This is the 'Let's go for it!' system.

(3) The Behavioural Inhibition System (BIS) is the most changed system in revised RST. It is responsible, not, as in the 1982 version, for mediating reactions to conditioned aversive stimuli and the special class of innate fear stimuli, but rather 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 conflicts; see Corr 2008a). In typical animal learning situations, BIS outputs have evolved to permit an animal to enter a dangerous situation (i.e., leading to cautious 'risk assessment' behaviour) or to withhold entrance (i.e., passive avoidance).

The BIS is involved in the processes that finally generate the emotion of anxiety, and 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, which is experienced subjectively as worry, apprehension and the feeling that actions may lead to a bad outcome; there is also an exaggerated startle reaction (Caseras, Fullana, Riba *et al.* 2006). The revised BIS resolves goal conflicts by increasing, through recursive loops, the negative valence of stimuli, via activation of the FFFS, until resolution occurs either in favour of approach or avoidance. In this important sense, there is a close relationship between the BIS and FFFS (see McNaughton and Corr 2008a).

The associated personality factor consists of worry-proneness and anxious rumination, leading to being constantly on the look-out for possible signs of danger, which map clinically onto such conditions as generalized anxiety and Obsessional-Compulsive Disorder (OCD). This is the 'Watch out, be very careful!' system. When activated by conflict stimuli, it is said to be in 'control mode', and when not activated, in 'just checking' mode (see Gray 1981). In support of this claim, using fMRI in a conflict paradigm, Haas, Omura, Constable and Canli (2007) found that the anxiety component of general Neuroticism was related to activation in the amygdala (see below).

Neural systems of FFFS-fear and BIS-anxiety

One major alteration in revised RST is the inclusion of a hierarchical arrangement of *distributed* brain systems that mediate specific defensive behaviours associated with level of threat experienced, ranging from the prefrontal cortex, at the highest level, to the periaqueductal grey, at the lowest level. To each structure is assigned a specific class of mental disorder (McNaughton and Corr 2008a). The evolution of these separate systems that form a whole system most probably evolved by a 'rule of thumb' (ROT) approach (McNaughton and Corr in press). According to this perspective, separate emotions (e.g., fear, panic, etc.) may be seen as reflecting the evolution of specific neural modules to deal with specific environmental demands (e.g., flee in the face of a predator) and, as these separate systems evolved and started to work together, some form of regulatory process (e.g., when one module is active, others are inactivated) evolved. The resulting hierarchical nature of this defence system reflects the fact that simpler systems must have evolved before more complex ones, which provides a solution to the problem of conflicting action systems: the later systems evolved to have inhibitory control on lower-level systems. The result of this process of evolution is the existence of hierarchically ordered series of defensive reactions, each appropriate for a given defensive distance (i.e., level of threat perceived; see below).

This hierarchical arrangement may seem at first to be complex; however, it can be conveniently summarized in terms of a two-dimensional scheme, consisting of 'defensive distance' and 'defensive direction' – the prize we win from tolerating some modicum of complexity is synthesis of a vast literature of research findings into a coherent whole, showing, for example, why psychological disorders have specific elements while at the same time showing co-morbidity with other disorders. The two-dimensional neural (CNS) theory translates this two-dimensional (cns) psychological schema, reflecting two broad affective dimensions (Figure 21.4).

We now turn to the two dimensions of this hierarchical neural arrangement: *defensive direction* and *defensive distance*.

Defensive direction: fear versus anxiety

The avoidance of, or approach to, a dangerous stimulus is reflected in the categorical dimension of 'defensive direction', which further reflects a functional distinction between behaviours (a) that remove an animal from a source of danger (FFFS-mediated, fear), and (b) that allow it cautiously to approach a source of potential danger (BIS-mediated, anxiety). These functions are ethologically and pharmacologically distinct and, on each of these separate grounds, can be identified with fear and anxiety, respectively. To better understand this distinction, a few words must be spent on the influential work of Robert and Caroline Blanchard (Blanchard and Blanchard 1988, 1990; Blanchard, Griebel, Henrie and Blanchard 1997), who were most responsible for moving Gray away from a formal analysis

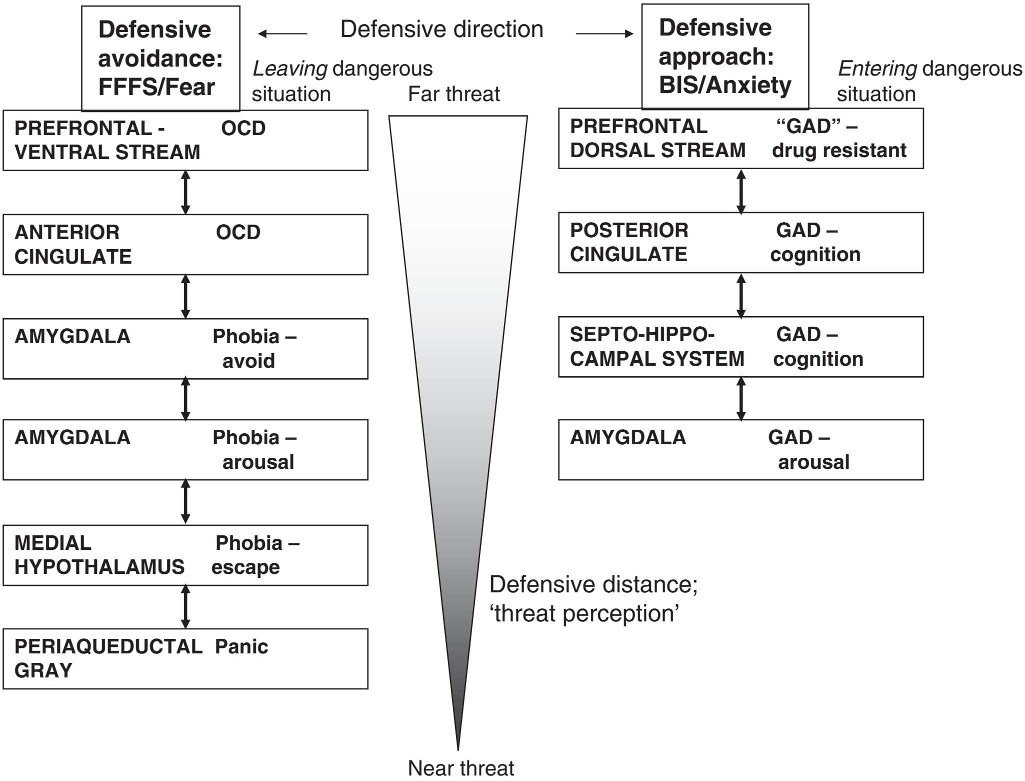
Defensive avoidance: FFFS/Fear
Defensive direction
Defensive approach: BIS/Anxiety
Leaving dangerous situation
Far threat
Entering dangerous situation
PREFRONTAL - VENTRAL STREAM OCD
ANTERIOR CINGULATE OCD
AMYGDALA Phobia – avoid
AMYGDALA Phobia – arousal
MEDIAL HYPOTHALAMUS Phobia – escape
PERIAQUEDUCTAL GRAY Panic
PREFRONTAL DORSAL STREAM "GAD" – drug resistant
POSTERIOR CINGULATE GAD – cognition
SEPTO-HIPPO-CAMPAL SYSTEM GAD – cognition
AMYGDALA GAD – arousal
Defensive distance; 'threat perception'
Near threat

Figure 21.4. *The two dimensional defence system. On either side are defensive avoidance and defensive approach, respectively (this is a categorical dimension of 'defensive direction'). Each is divided, down the page, into a number of hierarchical levels, both with respect to neural level (and cytoarchitectonic complexity) and to functional level (this is a qualitative dimension of 'defensive distance', or more generally 'threat perception'). Each level is associated with specific classes of behaviour and so symptom and syndrome (as shown).*

of behaviour based on learning theory (Gray 1975, 1982) to one based on functional classes of behaviour (e.g., freezing vs. cautious approach) (Gray and McNaughton 2000).

Over an extensive period of research, the Blanchards examined the behavioural effects of classes of psychiatric drugs on defensive behaviours of rodents in realistic experimental situations, known as 'ethoexperimental analysis': 'etho' to reflect the natural behaviours shown by rodents in real-like environments (e.g., freezing in the face of threat), and 'experimental' to reflect the control over the features of this reality-like environment (e.g., smell vs. presence of cat in the reality-like visual burrow designed by the Blanchards): to the rodents, this world is real enough, the threat stimuli are highly salient, and the behaviours observed and measured are not predefined by the experimenter (as would be the case with the use of a Skinner box). Careful analysis of the behavioural effects on rodents of clinically effective psychiatric drugs (e.g., anxiolytics) revealed a set of findings that pointed to the existence of two broad classes of defensive behaviour (avoidance of threat and cautious approach to threat) – or, in the Blanchards' view, immediate vs. potential threat. In passing, we should note that the Blanchards' research approach very much parallelled Gray's own (see above), therefore it is not surprising that their results were to prove of such value to Gray, along with colleague Neil McNaughton, in revising RST.

The Blanchards' findings may be summarized as follows. First, one class of behaviours was elicited by the immediate presence of a predator (e.g., a cat) – this class could clearly be attributed to a state of fear. The behaviours were observed to be highly sensitive to panicolytic (i.e., panic-reducing) drugs, but not so much to drugs that are specifically anxiolytic (i.e., anxiety-reducing). Secondly, a quite distinct class of behaviours (including 'risk assessment') was elicited by the potential presence of a predator – this class of behaviours was highly sensitive to anxiolytic drugs. Both functionally and pharmacologically, this class was distinct from the behaviours attributed to fear and could be attributed to a state of anxiety. As this distinction shows, in some important functional respects, fear and anxiety can reflect opposing motivations (avoiding vs. entering dangerous situations).

Defensive distance: fear and anxiety

The type of behavioural reaction to a threat is reflected in the second dimension of 'defensive distance', which reflects further the actual, or perceived, distance from threat. This dimension applies equally to fear and anxiety but operates differently in each case: anxiolytic drugs change it in the case of the BIS-anxiety, but not in the case of FFFS-fear. The main point is that defensive distance (i.e., how far you think you are from the threat, which closes with increasing magnitude of threat) corresponds to activation of specific neural modules (e.g., at very close defensive distance, PAG activation and panic): the common expletive 'Oh shit!' is more than being merely figurative, because one of the most reliable signs of intense fear in rodents and man (e.g., soldiers in battle) is defecation (Stouffer *et al*. 1950).

Although we can equate defensive distance with real distance, it is more accurately seen as a *perception*; that is, an internal quantity that defines defensive reactions to a fixed unit of threat (i.e., magnitude x distance). This rather humble statement provides an immediate explanation for 'neurosis'; that is, individual differences in the susceptibility to neurotic disorder. As shown in Table 21.1, a more defensive person (for simplicity here, defined so as to cut across both fear and anxiety) will *perceive* a threat of a fixed objective value as being more threatening (i.e., closer) than a less defensive person. Indeed, this hypothesis helps to explain the actions of drugs: they do not affect the *intensity* of a particular behaviour (e.g., avoidance); rather they affect 'perceived distance' (i.e., the magnitude of perceived threat), and thus they lead to *different* behaviours being shown (e.g., from avoidance to cautious approach).

Table 21.1 *Relationship between personality trait of 'defensiveness' (FFFS/ BIS), difference between actual and perceived defensive distance, and the real defensive difference required to elicit defensive behaviour.*

Personality trait	Defensive distance	Real defensive distance required for elicitation of defensive behaviour
High defensive individual	Perceived distance < actual distance	Long
Normal defensive individual	Perceived distance = actual distance	Medium
Low defensive individual	Perceived distance > actual distance	Short

This form of analysis counsels us not to focus on behaviour per se, but rather to view behaviour as a reflection of central states of emotion and motivation: as an overt, and measurable, indicator of internal states. Much of behavioural pharmacology results would simply not make sense if we only looked at the intensity of a particular behaviour. This point deserves emphasizing. Take a foraging (conflict) situation in which the perceived intensity of threat is high (i.e., small defensive distance). An animal that is not drugged is likely to remain behaviourally still and anxiolytic drugs serve to *increase* risk assessment (i.e., lead to behavioural exploration). However, if the perceived threat is only medium, now the undrugged animal is likely to engage in exploratory, risk-assessment behaviour and anxiolytic drugs will serve to *decrease* risk-assessment behaviour (because the animal is now experiencing the threat as more distant and is no longer anxious and, thus, returns to normal appetitive behaviour). The important point is that the drug does not alter a specific risk assessment in any simple fashion, but leads to changes in behaviour that depend on the animal's internal state (Blanchard and Blanchard 1990).

BIS-mediated conflict

As noted above, the BIS has been substantially revised and updated: it is now defined in terms of defensive *approach* (i.e., behavioural caution in a rewarding environment, e.g., foraging). However, revised RST argues that this behaviour, along with the previously emphasized conditioned aversive stimuli that were said to activate the BIS, are only examples of a more fundamental aspect of the BIS, namely that it is sensitive to goal *conflict* (e.g., approach-avoidance; e.g., an animal will approach a threat only if there is some possibility of a rewarding outcome, such as food). However, threats (as opposed to primary punishment itself) are only one source of aversion. Revised RST argues that, in principle, approach-approach and avoidance-avoidance conflicts also involve activation of the same system and have essentially the same effects as the classic approach-avoidance. An example of an approach-approach conflict is: which equally appealing job should you take? The aversive element resides in the possibility of making a mistake, thus we typically spend time weighing up all the possibilities, and searching for potential downsides to each decision. We may speculate – and it can only be that – that much of modern-day angst comes from the conflicting choices available in our successful economic system. Novelty is another type of stimuli that may activate the BIS (although, if sufficiently intense it is likely to activate the FFFS) as it entails a conflict between what is expected and what is perceived. Little research attention work has been devoted to this aspect of BIS theory, however one study has provided evidence for a preference for familiarity (as opposed to novelty) in high BIS individuals (Quilty, Oakman and Farolden 2007).

Before ending this section, an important asymmetry must be noted: fear can be generated without a significant degree of anxiety (i.e., in the absence of goal-conflict), but BIS activation always leads to FFFS activation via the increase in negative valence. For this reason FFFS and BIS will often be co-activated – and, as we will see below, this is a good reason for lumping them together into a single 'Punishment Sensitivity' factor of personality.

This revised view of the BIS is starting to explain previous anomalies in the literature and is pointing to new research questions. For example, Wallace and Newman (2008) discussed the relationship between an impaired BIS and psychopathy, which was in the old version of the theory associated with an absence of anxiety (and fear more broadly). However, these authors note that the evidence in favour of an impairment of anxiety/fear in psychopaths is weak; indeed, under certain conditions, psychopaths display normal reactions when anxiety/fear is present. Wallace and Newman (2008) point to the response modulation deficit seen in psychopathy, which impairs responses to aversive stimuli when a dominant response set to reward has been established. The revised conception of the BIS explains this finding: an impaired BIS does not signal prepotent (BAS-related) response conflict when environmental contingencies change to favour aversive motivation and avoidance, and in consequence the psychopath does not respond in an adaptive manner to the presence of aversive stimuli. BIS underactivity seems to be especially marked in primary (low fear) psychopathy (Ross, Mottó, Poy *et al.* 2007).

Behavioural approach system

There is little new to add on the BAS in terms of the Gray and McNaughton (2000) revision. However, work by the author, as well other RST researchers (e.g., Pickering 2008), have highlighted a number of issues that require attention. One such issue concerns the complexity of the BAS and the implications of this complexity for personality measurement. Elsewhere (Corr 2008a), I have pointed out that, on evolutionary grounds, it may be assumed that the BAS is more complex than conventionally thought – and, indeed, may be more complex than either the FFFS or the BIS.[6] I (Corr 2008a) developed the concept of *sub-goal scaffolding*, which reflects the separate, though overlapping, stages of BAS behaviour, consisting in a series of appetitively-motivated sub-goals. Sub-goal scaffolding reflects the fact that, in order to move along the temporo-spatial gradient to the final primary biological reinforcer, it is necessary to engage a number of distinct processes. Complex approach behaviour entails a series of behavioural processes, some of which oppose each other. Such behaviour often demands *restraint* and *planning*, but, especially at the final point of *capture* of the biological reinforcer, impulsivity is more appropriate. Therefore, simply being a highly impulsive person (i.e., not planning and acting fast without thinking) would be detrimental to effective BAS behaviour. For these reasons, 'impulsivity' may not be the most appropriate name for the personality dimension that reflects BAS processes (Franken and Muris 2006; Smillie, Jackson and Dalgleish 2006)

There is evidence that, at the psychometric level, the BAS is multidimensional. For example, the Carver and White (1994) BIS/BAS scales measure three aspects of BAS: Reward Responsiveness, Drive and Fun-Seeking – these scales have good psychometric properties in both adolescents and adults (e.g., Caci, Deschaux and Baylé 2007; Cooper, Gomez and Aucute 2007). In accordance with the concept of sub-goal scaffolding, we may see that Drive is concerned with actively pursuing desired goals, Reward-Responsiveness is concerned with excitement at doing things well and winning, and Fun-Seeking is concerned with the impulsivity aspect of the BAS.

There is also the issue of the involvement of the BAS in negative emotional states. On the basis of an analysis of the BAS and frustrative non-reward, it has been hypothesized that reward sensitive individuals would be the first to detect a lower than expected level of reward and, thus, experience frustration (Corr 2002b; see also Carver 2004, and Harmon-Jones 2003). Important in this regard is the system that mediates these negative states: must it be either the FFFS or the BIS, or might only the BAS be involved, and if the latter, how?

Personality factors

So far we have equated 'personality' with individual variations in the major brain-behavioural systems that underlie the FFFS, BAS and BIS. Existing RST

[6] The 'life-dinner principle' (Dawkins and Krebs 1979) suggests that the evolutionary selective pressures on prey are much stronger than on predators: if a predator fails to kill its prey then it 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 behaviours (e.g., freezing, fleeing and defensive attack) are 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, which entail a higher degree of organization and planning. In addition, the heterogeneity of appetitive goals (e.g., securing food and finding/keeping a sexual mate) demands a heterogeneity of BAS-related strategies: no one set of behaviours would be sufficient to achieve these very different BAS goals.

questionnaire measures were developed on the basis of the pre-2000 theory. For example, in addition to the three sub-scales of the Carver and White (1994) BAS scale, it provides an apparently unitary measure of BIS. Importantly, however, fear and anxiety are not differentiated. To some extent, within the BIS scale it is possible to separate fear from anxiety (Corr and McNaughton 2008; putative FFFS-Fear and BIS-Anxiety in square brackets), although for some items this differentiation is blurred.

(1) Even if something bad is about to happen to me, I rarely experience fear or nervousness. [FFFS]
(2) Criticism or scolding hurts me a lot. [FFFS/BIS]
(3) I feel pretty worried or upset when I think or know somebody is angry at me. [FFFS/BIS]
(4) If I think something unpleasant is going to happen I usually get pretty 'worked up'. [FFFS/BIS]
(5) I feel worried when I think I have done poorly at something. [BIS]
(6) I have few fears compared to my friends. [FFFS]
(7) I worry about making mistakes. [BIS]

Poythress, Skeem, Weir *et al.* (2008) reported that, in an offender sample, the BIS scale does, indeed, break down into two sub-scales, as indicated above (see also, Johnson, Turner and Iwata 2004), suggesting that closer attention should be paid to differentiating fear and anxiety even in existing questionnaires. However, if we are interested in measuring non-specific punishment sensitivity then a conflation of FFFS-fear and BIS-anxiety may work quite well, and this possibility may account for the popularity of the BIS scale of the Carver and White scales. In terms of revised RST, Corr and McNaughton (2008) inclined to the view that the old 'Anxiety' axis (i.e., Neurotism-Introversion) reflects 'Punishment Sensitivity', or 'Threat Perception', or simply 'Defensive Distance', with lower-order factors of this orthogonal 'dimension' breaking down into specific oblique FFFS-fear and BIS-anxiety factors. There remains much work needed to develop revised RST scales that display theoretical fidelity and psychometrical rigour. That the differentiation of fear and anxiety is needed in terms of personality scales is shown by the following studies. Recent structural equation modelling has confirmed the fear-anxiety differentiation hypothesis (Cooper, Perkins and Corr 2007), as have predictive validity studies (Perkins, Kemp and Corr 2007).

Cutting across the BAS, FFFS and BIS is physiological arousal – here we return to the main concern of Eysenck's theory. Concurrent activation of the FFFS, BIS and BAS sums in the production of general arousal; this summation of 'intensity' function, as distinct from the 'direction of behaviour', has a long history in behavioural psychology (e.g., Duffy 1962). This common summation of input from all the systems provides a source for a very general factor of 'arousability' that reflects changes in the responsiveness of the autonomic nervous system. We only now have to assume that Eysenck's Extraversion factor reflects the balance of reward and punishment systems (a central assumption in RST) for a viable explanation as to why Extraversion and arousal are so often associated in experimental studies of personality. So too, we might infer a general factor of emotional activation, reflecting the summed activity of reward and punishment systems, to derive a general dimension of Neuroticism.

We, thus, have a choice of personality levels of description. On the one hand, if we want to measure separable causally-efficient systems in the brain (i.e., FFFS, BIS and BAS), then we should opt for specific personality questionnaires that faithfully measure the activity of these systems. On the other hand, if we want to measure the net product of the interplay of these systems, then we should opt for Eysenckian-type personality questionnaires that measure broad dimensions of personality (e.g., Extraversion and Neuroticism) relating to broad neurophysiological factors (e.g., arousal). We may further want to measure, in addition to these factors, those relating to styles of personality (e.g., Agreeableness in the Five-Factor Model). Each of these levels of analysis are complementary.

Personality and psychopathology

The two constructs of 'defensive direction' and 'defensive distance', and their mapping onto the series of neural modules that comprise the FFFS and BIS which, in turn, are attributed particular functions, can be related to common symptomatology (see Figure 21.5).

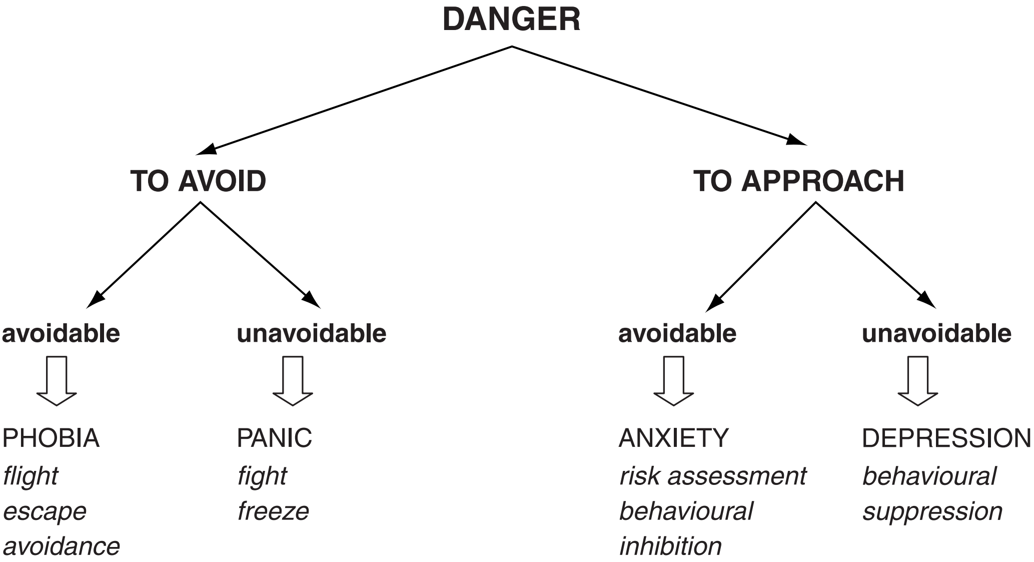

Figure 21.5. *Categories of emotion and defensive responses derived from 'defensive direction' (i.e., motivation to avoid or approach the source of danger) and avoidability of the threat (given constraints of the environment). Emotions in capitalization are psychiatric-based, and defensive behaviours in italics are derived from animal learning paradigms.*

In addition to hypersensitivity in a particular neural module giving rise to a specific set of symptoms (e.g., periaqueductual grey and panic), there are interactions of the FFFS and BIS that have important implications for explicating the underlying basis of a specific disorder. For example, pathologically excessive (BIS) anxiety could generate (FFFS) panic with the latter being entirely appropriate to the level of apprehension experienced. Also, pathological panic could, with repeated experience, condition anxiety with the level of the latter being appropriate to the panic experienced. This state of affairs means that symptoms alone may offer a misleading picture of the basic neural dysfunction. Specifically, hypersensitivity and activity in one neural module may well activate other modules as a secondary consequence and, furthermore, over time sensitize the whole defensive system to ease of activation. This may well explain the considerable comorbidity seen in neurotic conditions.

In a quite separate part of the psychopathology literature, the distinction between fear and anxiety has been identified. A behavioural genetic study of ten major psychiatric disorders, in a sample of 5,600 twins (Kendler, Prescott, Myers and Neale 2003) revealed the following findings: (a) two major dimensions emerged, one relating to *internalizing* disorders (i.e., major depression, generalized anxiety disorder and phobia), the other to *externalizing* disorders (i.e., alcohol dependence, drug abuse/dependence, adult antisocial behaviour and conduct disorder); (b) no differences in genetic and environmental influences for males and females, despite the large difference in prevalence rates; (c) unique (i.e., non-shared family) environment effects for internalizing disorders; (d) and, of most relevance to RST, the structure of genetic risk for internalizing disorders broken down into an 'anxious-misery' factor (i.e., depression, generalized disorder and panic) and a specific 'fear' factor (i.e., animal and situational phobia).

Earlier, Prescott and Kendler (1998) noted that mild depression and generalized anxiety do not appear to have distinct genetic etiologies, but rather a common genetic basis, perhaps a disposition to dysphoric mood which is shaped by individual experiences into symptoms of depression, anxiety, or both. (See also, Kendler *et al.* (1992.) As Kender *et al.* (2003, p. 935) themselves speculated, 'It is tempting to speculate that these genetic factors on risk might be mediated through personality.'

Indeed, this genetic risk structure for internalizing disorders – with one major factor breaking down into fear and anxiety sub-factors – is the same as that proposed in Figure 21.3 (here fear and anxiety factors are collapsed together to give a general punishment factor). Behavioural studies of rodent defensive behaviour are also starting to differentiate fear and anxiety (e.g., Tsetsenis, Ma, Iacono *et al.* 2007); this study also suggests that the hippocampus is important in the response to ambiguous aversive stimuli.

Important in this regard are quantitative genetic analysis of both change and continuity in BIS/BAS sensitivity over a period of two to three years. One study showed the following: genetic factors accounted for approximately one-third of variance in BIS and BAS; genetic factors contributed to continuity, but not change, whereas environmental factors accounted for both continuity and change in both traits. In this study, the degree of genetic influence did not differ across time (Takashasi, Ma, Iacono *et al*. 2007). On the basis of the relative magnitude of effects, these authors concluded that, at least in this age group (mean age early to mid-twenties), temporal stability of individual differences in these RST traits 'owes more to genetic than to environmental factors'. Given that the Carver and White BIS/BAS scales were used in this study, it would have been interesting if FFFS-fear and BIS-anxiety item clusters had been analysed separately.

Conclusions

Over a forty-year period, RST has developed into a sophisticated model of emotion, motivation, personality and psychopathology, and to this achievement we owe a debt of gratitude to the fundamental work of Jeffrey Gray. Although in a continual state of development, the general model of RST synthesizes vast literatures (e.g., behavioural pharmacology of emotion, motivation and learning) and forges bridges between hitherto unrelated areas (e.g., ethoexperimental studies and personality). Of importance is the translational nature of this research: we can now go from basic non-human animal studies to human ones, armed with a rigorous theory to guide the difficult process of understanding the neuropsychology of human personality. As an example of such translational research, Perkins and Corr (2006) confirmed that the basic defensive reactions of rodents to cats in ethologically-valid situations are found in human defensive reactions to a range of threatening situations.

There are many problems still to be addressed in RST, including the following (non-exhaustive) list: (a) how best to characterize BAS processes and how to measure them by questionnaire (Corr 2008a; Pickering and Smillie 2008); (b) what is the relationship between conscious awareness, its functions and emotion/motivation (Gray 2004; Corr 2006, 2008a); (c) how best to operationalize reward and punishment variables in the laboratory and what predictions we should make about their possible interaction (Corr 2002a, 2008a); (d) what is the most appropriate way to measure FFFS, BIS and BAS in human beings, and how such measures can be validated; and (e) are the principles of frustrative non-reward and relief of non-punishment useful in explaining counter-productive and paradoxical behaviour (McNaughton and Corr in press). RST may also have gone some way to help explain the phenomological nature of fear, anxiety and hope: why they 'feel' the way they do; however, it will be some time before we have a consensual model of why emotions are conscious in the first place – although, arguably, Gray (2004) himself has gone a long way to elucidating the functions of consciousness (Corr 2006, 2008a). On top of these problems are wider ones, ranging from the role of 'free will' in behaviour, and how individual behaviour is regulated by society (e.g., effective penal systems).

RST has come a long way, but it still has a long way to go before it can be said to provide a comprehensive model of emotion, motivation, personality and psychopathology. As shown in this chapter, it is a *general* theory that aspires to encapsulate most of the biologically-relevant findings, as well as having the capacity to incorporate new developments. Inevitably, the *specific* form of the theory, at any one ‘flash-bulb’ moment, will appear in certain respects ill-specified and incomplete.

This chapter has covered a lot of ground and encountered some of the difficulties and unresolved issues that remain; and it has revealed that we must continue to tolerate considerable uncertainty as to the best way to relate fundamental systems of emotion and motivation to personality factors and psychopathology – this is not unique to the RST but to the field in general. Although much work lies ahead, arguably, large areas of hitherto wild growth have been cleared away to reveal the fundamental terrain of the neuropsychology of personality.

References

Blanchard, D. C. and Blanchard, R. J. 1988. Ethoexperimental approaches to the biology of emotion, *Annual Review of Psychology* 39: 43–68

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

Blanchard, R. J., Griebel, G., Henrie, J. A. and Blanchard, D. C. 1997. Differentiation of anxiolytic and panicolytic drugs by effects on rat and mouse defense test batteries, *Neuroscience and Biobehavioral Reviews* 21: 783–9

Caci, H., Deschaux, O. and Baylé, F. J. 2007. Psychometric properties of the French versions of the BIS/BAS and the SPSRQ, *Personality and Individual Differences* 42: 987–98

Carver, C. S. 2004. Negative affects deriving from the Behavioral Approach System, *Emotion* 41: 3–22

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: 319–33

Caseras, F. X., Fullana, M. A., Riba, J., Barbanoj, M. J., Aluja A. and Torrubia, R. 2006. Influence of individual differences in the Behavioural Inhibition System and stimulus content (fear versus blood-disgust) on affective startle reflect modulation, *Biological Psychology* 72: 251–6

Cooper, A., Gomez, R. and Aucute, H. 2007. The Behavioural Inhibition System and Behavioural Approach System (BIS/BAS) scales: measurement and structural invariance across adults and adolescents, *Personality and Individual Differences* 43: 295–305

Cooper, A. J., Perkins, A. and Corr, P. J. 2007. A confirmatory factor analytic study of anxiety, fear and Behavioural Inhibition System measures, *Journal of Individual Differences* 28: 179–87

Corr, P. J. 2001. Testing problems in J. A. Gray’s personality theory: a commentary on Matthews and Gilliland (1999), *Personal Individual Differences* 30: 333–52

2002a. J. A. Gray's reinforcement sensitivity theory: tests of the joint subsystem hypothesis of anxiety and impulsivity, *Personality and Individual Differences* 33: 511–32

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: 1247–53

2004. Reinforcement sensitivity theory and personality, *Neuroscience and Biobehavioral Reviews* 28: 317–32

2006. *Understanding biological psychology.* Oxford: Blackwell

2007. Personality and psychology: Hans Eysenck's unifying themes, *The Psychologist* 20: 666–9

2008a. Reinforcement sensitivity theory (RST): Introduction, in P. J. Corr (ed). *The reinforcement sensitivity theory of personality*, pp. 1–43. Cambridge University Press

2008b. *The Reinforcement Sensitivity Theory of Personality*, Cambridge University Press

Corr, P. J. and McNaughton, N. 2008. Reinforcement sensitivity theory and personality, in P. J. Corr (ed). *The reinforcement sensitivity theory of personality*, pp. 155–87. Cambridge University Press

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

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

Duffy, E. 1962. *Activation and behaviour.* London: Wiley

Eilam, D. 2005. Die hard: a blend of freezing and fleeing as a dynamic defense – implications for the control of defensive behavior, *Neuroscience and Biobehavioral Reviews* 29: 1181–91

Eysenck, H. J. 1947. *Dimensions of personality.* London: K. Paul/Trench Trubner

1957. *The dynamics of anxiety and hysteria.* New York: Preger

1967. *The biological basis of personality.* Springfield, IL: Thomas

1979. The conditioning model of neurosis, *Behavioural and Brain Sciences* 2: 155–99

Eysenck, H. J. and Eysenck, S. G. B. 1976. *Psychoticism as a dimension of personality.* London: Hodder and Stoughton

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), *The biological bases of individual behaviour*, pp. 206–20. London: Academic Press

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*, pp. 7–34. New York: Guilford Press

Franken, I. H. A. and Muris, P. 2006. Gray's impulsivity dimension: a distinction between Reward Sensitivity versus Rash Impulsiveness, *Personality and Individual Differences* 40: 1337–47

Gray, J. A. 1964. *Pavlov's typology.* Oxford: Pergamon Press

1970. The psychophysiological basis of Introversion–Extraversion, *Behaivour Research and Therapy* 8: 249–66

1972a. Learning theory, the conceptual nervous system and personality, in V. D. Nebylitsyn and J. A. Gray (eds.), *The biological bases of individual behaviour*, pp. 372–99. New York: Academic Press

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*, pp. 182–205. New York: Academic Press

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*, pp. 3–41. London: Wiley

1977. Drug effects on fear and frustration: possible limbic site of action of minor tranquillizers, in L. L. Iversen, S. D. Iversen and S. H. Snyder (eds), *Handbook of psychopharmacology,* vol. VIII*, Drugs, neurotransmitters, and behavior*, pp. 433–529. New York: Plenum Press

1981. A critique of Eysenck's theory of personality, in H. J. Eysenck (ed.), *A model for personality*, pp. 246–76. Berlin: Springer

1982. *The neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system*. Oxford University Press

2004. *Consciousness: creeping up on the Hard Problem*. Oxford University Press

Gray, J. A. and McNaughton, N. 2000. *The neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system*. Oxford University Press

Haas, B. W., Omura, K., Constable, R. T. and Canli, T. 2007. Emotional conflict and neuroticism: personality-dependent activation in the amygdala and subgenual anterior cingulated, *Behavioural Neuroscience* 121: 249–56

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

Hebb, D. O. 1955. Drives and the C. N. S. (Conceptual Nervous System), *Psychological Review* 62: 243–54

Hull, C. L. 1952. *A behaviour system*. New Haven: Yale University Press

Johnson, S. L., Turner, R. J and Iwata, N. 2004. BIS/BAS levels and psychiatric disorder: an epidemiological study, *Journal of Psychopathology and Behavioral Assessment* 25: 25–36

Kendler, K. S., Neale, M. C., Kessler, R. C., Heath, A. C. and Eaves, L. J. 1992. Major depression and generalized anxiety disorder: same genes, (partly) different environments?, *Archives of General Psychiatry* 49: 716–22

Kendler, K. S., Prescott, C. A., Myers, J. and Neale, M. C. 2003. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women, *Archives of General Psychiatry* 60: 929–37

Konorski, J. 1967. *Integrative activity of the brain*. 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*, pp. 91–196. Cambridge University Press

Lykken, D. T. 1971. Multiple factor analysis and personality research, *Journal of Research in Personality* 5: 161–70

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

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

2008a. The neuropsychology of fear and anxiety: a foundation for reinforcement sensitivity theory, in P. J. Corr (ed). *The reinforcement sensitivity theory of personality*, pp. 44–94. Cambridge University Press

2008b. Animal cognition and human personality, in P. J. Corr (ed.), *The Reinforcement Sensitivity Theory of Personality*, pp. 95–119. Cambridge University Press

in press. Central theories of motivation and emotion, in G. G. Berntson and J. T. Cacioppo (eds), *Handbook of neuroscience for the behavioural sciences*. London: Wiley

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: 419–27

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: 21–8

Perkins, A. M., Kemp, S. E. and Corr, P. J. 2007. Fear and anxiety as separable emotions: an investigation of the revised reinforcement sensitivity theory of personality, *Emotion* 7: 252–61

Pickering, A. D. 2008. Format and computational models of Reinforcement Sensitivity Theory, in P. J. Corr (ed.), *The Reinforcement Sensitivity Theory of Personality*, pp. 453–81. Cambridge University Press

Pickering, A. D., Díaz, A. and Gray, J. A. 1995. Personality and reinforcement: an exploration using a maze-learning task, *Personality and Individual Differences* 18: 541–58

Pickering, A. D. and Smillie, L. D. 2008. The behavioural activation system: challenges and opportunities, in P. J. Corr (ed). *The reinforcement sensitivity theory of personality*, pp. 120–54. Cambridge University Press

Poythress, N. G., Skeem, J. L., Weir, J., Lilienfeld, S. O., Douglas, K. D., Edens, J. F. P. and Kennealy, J. 2008. Psychometric properties of Carver and White's (1994) BIS/BAS scales in a large sample of offenders, *Personality and Individual Differences* 45: 732–7

Prescott. C. A. and Kendler, K. S. 1998. Do anxious and depressive states share common genetic factors?, *European Neuropsychopharmacology* 8 Suppl. 2: S76–S77

Quilty, L. C., Oakman, J. M. and Farolden, P. 2007. Behavioural inhibition, behavioural activation, and the preference for familiarity, *Personality and Individual Differences* 42: 291–303

Revelle, W. 1997. Extraversion and impulsivity: the lost dimension, in H. Nyborg (ed.), *The scientific study of human nature: tribute to Hans J. Eysenck at eighty*, pp. 189–212. Oxford: Elsevier Science Press

Ross, S. R., Mottó, J., Poy, R., Seganra, P., Pastor, M. C. and Montañés, S. 2007. Gray's model and psychopathy: BIS but not BAS differentiates primary from secondary psychopathy in noninstitutionalized young adults, *Personality and Individual Differences* 43: 1644–55

Skinner, B. F. 1953. *Science and human behaviour*. New York: Macmillan

Smillie, L. D., Jackson, C. J. and Dalgleish, L. I. 2006. Conceptual distinctions among Carver and White's (1994) BAS scales: a reward-reactivity versus trait impulsivity perspective, *Personality and Individual Differences* 40: 1039–50

Smillie, L. D., Pickering, A. D. and Jackson, C. J. 2006. The new reinforcement sensitivity theory: implications for personality measurement, *Personality and Social Psychology Review* 10: 320–35

Spence, K. W. 1964. Anxiety (drive) level and performance in eyelid conditioning, *Psychological Bulletin* 61: 129–39

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A. and Clausen, J. A. 1950. *Studies in social psychology in World War II,* Vol. IV*, Measurement and prediction*. Princeton University Press

Takashasi, Y., Yamagata, S., Kijima, N., Shigemasu, K., Ono, Y. and Ando, J. 2007. Continuity and change in behavioural inhibition and activation systems: a longitudinal behavioural genetic analysis, *Personality and Individual Differences* 43: 1616–23

Tsetsenis, T., Ma, X.-H., Iacono, L. L., Beck, S. G. and Gross, C. 2007. Suppression of conditioning to ambiguous cues by pharmacogenetic inhibition of the dentate gyrus, *Nature Neuroscience* 10: 896–902

Wallace, J. F. and Newman, J. P. 2008. Reinforcement Sensitivity Theory and psychopathy: associations between psychopathy and the behavioral activation and inhibition systems, in P. J. Corr (ed). *The reinforcement sensitivity theory of personality*, pp. 398–414. Cambridge University Press

Weiner, N. 1948. *Cybernetics, or control and communication in the animal and machine*. Cambridge: MIT Press