5 Reinforcement Sensitivity Theory and personality

Philip J. Corr and Neil McNaughton

Reinforcement Sensitivity Theory (RST) is composed of two main components: (a) a *state* description of neural systems and associated, relatively short-term, emotions and behaviours; and (b) a *trait* description of longer-term dispositions to such emotions and behaviours. McNaughton and Corr (chapter 2) outlined the state level of description; this chapter explores the trait level of description and takes a more general view of the problems posed by the revised Gray and McNaughton (2000) theory.

'Top-down' and 'bottom-up' approaches to personality

The standard biological approach to personality adopts the well-established procedure from biology: first describe (taxonomy) and then explain (theorize about form of taxonomy; e.g., evolution). As noted by Gray (1972a, p. 372), 'The study of personality is the attempt (a) to discover consistent patterns of individual differences and (b) to account for the form taken by these patterns'. This 'top-down' approach has considerable merit and many empirical successes to its name. But it cannot be applied in a simple one-step fashion. Even within species and genera, taxonomy ((*a*) above) is not independent of causal theories ((*b*) above) – findings in molecular biology can alter taxonomy based on superficial description. With the study of personality it is a moot point whether the underlying variation in sensitivity of causal brain systems – which *must* control the psychological phenomena we classify under 'personality' – correspond in any obvious fashion to the manifest aspects of personality (i.e., factors, traits, facets, etc.). This chapter draws out some of the implications for personality research of understanding the relationship between (a) causal systems and (b) behavioural expressions, in an attempt to clarify the problems that future RST research will need to tackle.

Defining the problem?

The problem of relating causal and descriptive systems of personality is best illustrated by the seminal work of Hans Eysenck. Starting in 1944 with a statistical classification of individual variation in medical checklist items of 700 'war neurotics', which led to the postulation of the factors of Extraversion (E) and Neuroticism (N), Eysenck went on to propose causal theories to account for the brain-behavioural bases of these differences (inhibition-excitation theory in 1957, and arousal theory in 1967). This approach was adopted by Gray (1970, 1972b), who accepted that factor analysis can identify the minimum number, and thus the necessary (albeit not necessarily sufficient) factor space of personality, but not the rotation of axes (and hence causal systems) within that factor space.

However, close attention to the details of biological systems suggest that the factor analytic approach provides a description of personality that, whilst valid at the level of behavioural expression, fails adequately and sufficiently to reflect separable causal influences: thus, it may not be possible to use existing structural models of personality as a useful guide to discovering underlying causal systems. The conclusion from the analysis developed in this chapter is that if the phenotypic description of personality is not anchored to known brain systems then it will provide factors that are ill-matched to the underlying (genotypic and ontogenetic) causal processes. Accordingly, the growth in understanding the biology of personality will be stunted.

It needs to be borne in mind that the current uncertainty as to the best way to relate fundamental systems of emotion and motivation to personality factors is not a flaw in RST, but part of its ongoing development process, the nature of which has been described by Smillie, Pickering and Jackson (2006). They note that although RST is most often seen as a theory of anxiety and impulsivity, 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' (p. 320); they go on to remark, 'RST did not develop as a theory *of* specific traits, but as a theory of specific biological systems which were later suggested to relate, inter alia, to personality' (p. 321). There is another reason why basic emotion and motivational systems do not map neatly onto personality factors: basic emotion and motivation theory has extended beyond the point at which Gray suggested that the BIS and BAS relate to anxiety and impulsivity, respectively; and RST researchers have developed scales to measure the BIS and BAS that were influenced by Gray's original thinking and which do not reflect more recent developments in the basic theory. Thus, RST research represents two distinct bodies of knowledge, the first concerned with neural processes, the second with personality measurement. The Janus-face of RST is a strength, making it a dynamically evolving theory, but it also poses obvious problems for, at any given time, specifying a consensual model agreed by researchers. In particular, '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 *et al.* 2006, p. 321).

A biological approach to factor analysis

The class of statistical reduction methods that fall under the rubric of 'factor analysis' have played a fundamental role in personality research. The number, but not nature, of the sources of variance can be estimated by Principal Components Analysis (PCA) or one of the various forms of common factor analysis (CFA). These statistical techniques have been criticized for allowing arbitrary decisions to be made concerning the number of factors to be extracted and the location of the factors within the factor space. A biological perspective shows that such apparently arbitrary choices concerning the relation of variables (including items in a questionnaire) are not the correct means of settling the matter of the structure of personality, either in terms of the location of the axes within the space or, indeed, of whether factor axes should be orthogonal (for a discussion of the limitations of factor analysis, especially of complex biological systems, see Lykken 1971).

A purely mathematical choice of which scales to use, based on some criterion such as the nominal factorial purity of a pair of scales, has nothing to do with where a real underlying causal factor is controlling variance within the data space. The choice of scales is, however, important if we wish them each to be pure measures of a real causal factor. It is merely a convenient initial simplification to use orthogonal axes and rotate these so that the first factor captures the maximum shared variance among all the measured variables. As we get more knowledge of underlying causal systems, then *this* knowledge should influence the description of personality. Nor does the identification of a single real factor (e.g., Extraversion) commit one to dependence of this factor on a single cause. Thus, genetic influences, developmental changes, infection, long-term social factors and some single event in the previous week may all be important in causing personality expression. Of course, there is no necessary conflict between factor analytically-derived personality factors and those suggested by biological theory.

Extraversion and Neuroticism may make a good job of accounting for systematic individual differences at the highest level of description; but they may be a poor starting point if we wish to understand the complex of underlying causal systems – but often a 'poor' starting point is better than none at all. What is needed is a dynamic descriptive model, showing all levels of the structural hierarchy and how each level relates to each other – but we need to understand the functioning of the underlying causal systems first.

It might seem that we have manoeuvred ourselves into a somewhat awkward position: we have argued that factor analysis, upon which most of the established models of personality are based, may not provide a reliable guide to the biological basis of personality, yet that is where RST started and continues to thrive: i.e., on factor analytically-derived Extraversion and Neuroticism and other psychometric models of reward and punishment sensitivities. But experimental analysis cannot proceed until at least some kind of descriptive framework is in place for personality characters to be explained. Personality psychology must, therefore, start with factors discoverable by factor analysis. But crucially it must also move beyond this technique in developing and refining its description of personality, undertaking a continual process of anchoring descriptive axes and extending the dimensions of the factor space already discovered.

We must, therefore, conclude that factor analysis provides only a preliminary guide to the biological processes underlying the most common sources of variation in a population. This conclusion is demanded by the fact that factor analysis will not be able to differentiate, for example, separate causes that are conflated in development, and nor is it able to identify primary causes. It works on measures of the phenotype that may be (and often are) the end product of a long chain of causal, and interacting, influences. Causal and phenotypic factors may be so similar as to allow a one-to-one correspondence, but this would be the outcome of serendipity not of the logic of factor analysis. Therefore, 'discovering' a nominally single factor of personality and then *assuming* that there is a single causal basis of that factor is, to our way of thinking, a flawed strategy in the neuroscience of personality. It is an open empirical question whether there exists a single causal factor for the recovered personality factor.

Extraversion/neuroticism and reward/punishment sensitivity

Let us now take a closer look at the personality factors in the context of RST. Gray and Eysenck both accept that factor analysis recovers, at least, two personality dimensions: Extraversion and Neuroticism. The critical issues in relation to the factors within such a two-dimensional space is whether they are independent of each other (orthogonal) or not, and where in the space they should be placed. But the item loadings on Eysenck's (1944, 1947) factor analysis of neurotic symptoms/behaviours show that a large number of decisions had to be made concerning the nature of the normal personality dimensions corresponding to the (four original) factors of the factor analysis.[1] The later choice of items for the measurement of Extraversion and Neuroticism scales was not based on consideration of causal factors; it was based on preconceptions (i.e., a theory) about the most appropriate scales to *describe* the phenotypic nature of personality seen in the light of then current psychiatric nosology and other theoretical accounts of personality (e.g., Jung's Introversion-Extraversion). Eysenck's model was a 'best-guess'; indeed, it should be acknowledged as 'excellent-guess' given the ubiquity of these factors in virtually every other structural model of personality. But, as we have already seen, a valid structural model of factors at the most general level of description (i.e., dimensions) does not necessarily inform investigation of the associated causally-efficacious underlying systems.

Psychometric refinement

Eysenck's structural model has seen a number of changes over the years, and this psychometric refinement cannot be quoted in strong support of the original model (Gray 1981). Corr and McNaughton (in preparation) discuss this matter at some length. Here it is sufficient to note that Eysenck's (e.g., Eysenck and Eysenck 1975) attempt to create factorial purity represents, in essence, an arbitrary choice of axis rotation: he could have added a large number of other items to his scales and then, provided the item population had retained dimensional purity, he could have created orthogonal scales of reward sensitivity and punishment sensitivity. RST researchers have done exactly this in their creation of scales of reward and punishment sensitivity. Psychometric tinkering can take us only so far in understanding the causal bases of personality – it leaves whole areas barren of either description or explanation. The substantive issue for RST is how far our current knowledge of the brain

[1] When the four factors of Eysenck's (1944) matrix are rotated in accordance with conventional techniques, then these factors could, equally as well, be interpreted as reward and punishment sensitivity (Perkins, Revelle and Corr, unpublished); it is important to note that E and N were hypothesized from unrotated factors in 1944.

can identify the location of real factor (causal) axes. It is only when this matter is decided that the issue of the best measurement scale need be addressed. McNaughton and Corr (chapter 2) shows much progress has been made in identifying putative brain systems underlying personality; i.e., the *Fight-Flight-Freeze System* (FFFS), the *Behavioural Inhibition System* (BIS) and the *Behavioural Approach System* (BAS).

Nature of the extracted factors

The rotation of Extraversion and Neuroticism by 30° suggested by Gray (1970, 1972b), is sufficiently small that, to a first approximation, 'Neuroticism' could be seen as composed largely of 'Sensitivity to punishment' (perhaps comprising both fear and anxiety). Independent of whether we choose to rotate the axes, there is a separate question of the psychological nature of the factors being measured. There are thus two problems to solve:

(1) the location of the axes: Eysenck may have appeared to resolve this issue for his theory by fiat – refining the Extraversion scale to become independent to Neuroticism – but, as we have seen, either an alternative conception of his 1944 factor matrix or subsequent refinement of scales would have produced scales reflecting an alternative rotation;
(2) underlying systems: the specific functional nature of the underlying factor giving rise to each dimension; in practice, solving this issue should also solve issue 1, but until this solution is definitely achieved both need to be assessed.

Neural systems and personality factors

The main issue to be faced is to what extent variation in the sensitivity of neural or hormonal modulatory systems does or could result in personality factors. Tightly linked to this is the question of how the neural systems (BAS, FFFS, BIS) are modulated. There are a number of quite distinct (but not mutually exclusive) possibilities.

(1) The simplest possibility is that a specific modulatory system could act solely on one single functional system and could act uniformly on all the elements of that system. With three separate modulatory systems, this would generate a separate personality factor for each of the BIS, FFFS and BAS. According to this view, a single personality factor would predict all behaviour mediated by the FFFS, irrespective of the specific

neural circuits mediating specific behaviours (e.g., phobia = hypothamalus + amgydala), and likewise for the BAS and BIS.

(2) Next in complexity is the possibility that a specific modulatory system could act on two systems and would act uniformly on all the elements of both systems. There could, for example, be a single personality factor representing the sensitivity of both the FFFS and BIS concurrently (perhaps identifiable with 'threat sensitivity') or of both the BAS and FFFS (perhaps identifiable with 'reinforcer sensitivity'). At the personality level, we would need to take into account the distinct nature of the single modulatory system and of the two functional systems modulated. Critically we would expect co-variation, at the personality level, between measures that, at the state level, were selective for each of the systems.

(3) Most general and complex is the possibility that a specific modulatory system could act on selected parts of one or more functional systems. At the personality level, we would need to take into account both the nature of the modulatory system and the detailed nature of its selective action on the functional systems affected.

When we look more closely at the detailed neurology of the Gray and McNaughton (2000) theory, the situation becomes clearer. Both the FFFS and the BIS are represented by very large numbers of interlinked brain structures. But this multiplicity reflects a hierarchical organization that selects particular behaviours to match particular 'defensive distances' (or threat perception; see McNaughton and Corr, chapter 2). Threat perception itself represents the dimension of defensive distance (i.e., actual or perceived distance from threat) that selects, from all of the levels of the system, the currently appropriate one.

For the operation of any general factor underlying sensitivity to threat there must, then, be some neural or hormonal system that can modulate all levels of the BIS and/or FFFS concurrently. There are a number of potential candidates for this role in the theory: noradrenergic input, serotonergic input (BIS and FFFS), the endogenous hormonal ligand of the benzodiazepine receptor (the latter is restricted to the BIS only) and the various hormones in the hypothalamus-pituitary-adrenal (HPA) axis stress cascade (BIS and FFFS). Each of these has the capacity to affect many structures, in the same way, in parallel. Both noradrenergic and serotonergic systems can do this because they have multiple divergent collaterals targeting many structures. An endogenous benzodiazepine ligand and the various stress hormones could do this because of the widespread distribution of the relevant receptors across the critical structures and the delivery of the critical compounds to those structures via the blood stream.

Finally, cutting across the BAS, FFFS and BIS, is physiological arousal. Concurrent activation of different motivational systems within each of the BAS and FFFS is held (Gray and Smith 1969) to sum in the production of arousal. Activation of the BIS also increases arousal directly via the amygdala in a way that by-passes the septo-hippocampal control of behavioural inhibition. This common summation of input from all the systems provides an obvious potential source for a very general factor of 'arousability' that reflects changes in the responsiveness of the autonomic nervous system that are significant for all three of the BAS, FFFS and BIS.

We have, then, a range of physical substrates in which long-term changes in sensitivity could give rise to personality factors.

(1) An effect general within the BIS but relatively specific to it could be produced by changes in benzodiazepine-like hormones or their receptors. Determination as to whether any particular trait measure (or factor score) reflected activity in this system would be straightforward: such a trait measure should be reduced by repeated administration of a benzodiazepine agonist over a period of at least two weeks. As an additional check, buspirone could be repeatedly administered: this drug has the same action on the BIS as a benzodiazepine but essentially opposite side-effects and the latter, again, diminish with repeated administration. Any scale or measure intended to relate to the BIS should have, as a primary criterion, that it be similarly affected by long-term administration of these two chemically distinct types of anxiolytic. Given such a relation, its specificity would then have to be determined.

(2) Effects restricted to parts of the BIS could be produced by changes in noradrenaline or serotonin. To determine whether any particular trait measure (or factor score) reflected activity in the monoamine systems one would use chronic administration of either serotonin or noradrenaline re-uptake inhibitors, or a combination of the two. Changes in both together can also be produced by monoamine oxidase inhibitors. Any scale or measure affected by such monoaminergic manipulations but not by both benzodiazepines and buspirone would not be assessing BIS activity.

(3) General effects could be produced by changes in autonomic reactivity or by changes in circulating hormones, such as adrenaline.

(4) There are insufficient data to be sure whether the FFFS could be specifically modulated. While there are drugs that are panicolytic these tend to also be anxiolytic. Genetic or long-term environmental effects on the serotonin system are likely to be general to the FFFS and BIS rather than specific to the FFFS. Given the special focus of BIS output on the

FFFS (as a means of increasing negative bias), it may not be reasonable to expect there to be any system that modulates the FFFS without also having separate direct effects on the BIS – however, the functional hierarchies of the FFFS and BIS are clearly different, and the existence of specific psychometric measures of fear and anxiety, which can be shown to be relatively uncorrelated, points to the existence of separate causal influences.

(5) Effects selective to the BAS can probably be produced by chronic alterations in dopaminergic and/or opiodergic systems.

Monoamine systems and personality

The monoamine systems have been implicated in anxiety and the clustering of anxiety disorders is the basis for Gray's suggestion that Eysenck's factors should be rotated. An immediate point to note is that serotonin and noradrenaline are affected in similar ways by stress. They also have effects that combine synergistically. It follows that many genetic sources controlling serotonin and noradrenaline, and especially those that control monoamine oxidase (which has parallel effects on serotonin and noradrenaline), could underlie a single recoverable personality dimension. This point is of particular relevance to the only animal model we have for a relevant personality dimension.

There is evidence that 'emotionality' in rats is a homologue of neurotic introversion, or perhaps 'trait anxiety' (Broadhurst 1960). It has a major genetic component (Hall 1951; Gray 1987). Selective breeding for high and low emotional defecation ('emotionality') has resulted in the Maudsley Reactive and Maudsley Non-reactive strains of rat, respectively. Such defecation in anticipation of an aversive event occurs in many species (Candland and Nagy 1969), including our own (Stouffer *et al.* 1969; cited by Broadhurst 1960). It appears to be an indicator of very high levels of fear or anxiety (Hunt and Otis 1953).

Despite being selected for a single, rather basic, character, the two Maudsley strains differ on a huge range of items (Broadhurst 1975; Blizard 1981; Gray 1987). Overall, these items suggest that Maudsley Reactive rats have a greater response to threats in general rather than to anxiety-provoking stimuli in particular. The two strains also differ in an animal model of depression (Abel, Altman and Commissaris 1992; Viglinskaya *et al.* 1995) and this effect shows a strong linkage between changes in scores on an anxiety test with those on the depression test (Commissaris *et al.* 1996). Just like human neurotic introversion, the genetic differences between Reactive and Non-reactive rats appear to influence susceptibility to the full spectrum of neurotic reactions.

Long-term changes in the monoamine systems appear, then, to have general effects on defensive distance (i.e., actual or perceived distance from threat) that are independent of defensive direction (i.e., to avoid or approach the threat). Both with the direct alteration of serotonergic function (Deakin and Graeff 1991) and with the indirect genetic alteration of both noradrenergic and serotonergic function in the Maudsley strains, there are changes in both defensive approach and defensive avoidance. As noted above, the breadth of monoamine effects is consistent with the very broad morbidity (ranging from obsessive compulsive disorder through neurotic depression) associated with genetic and environmental alterations in neurotic introversion and 'punishment sensitivity' in general.

However, although these effects clearly range across the FFFS and BIS, they do not appear to be a change in threat sensitivity, pure and simple. It has been emphasized by Deakin and Graeff (Graeff 1994; Deakin and Graeff 1991) that the serotonergic system has opposite effects on panic to other aspects of defence. The separation between components of threat extends further. Serotonergic dysfunction appears to impair fear-related behaviour that requires a specific decision, such as passive avoidance and conditioned suppression, but appears to increase fear responses related to arousal, such as fear potentiated startle and immediate responses to aversive stimulation (Gray and McNaughton 2000, table 5.2). Although the available evidence is slim, it suggests that neurotic-introversion, or what Gray (1970) labelled 'trait anxiety', personality results from genetic or environmental changes in the long-term tone of the serotonin and noradrenaline systems – probably interacting synergistically. It also suggests that such changes will involve a general increase in many fear-(FFFS) and anxiety-(BIS) related behaviours but, at least in the general population, could involve a modest decrease in the tendency to panic.

Benzodiazepine receptors, BIS and personality

There could also be a personality factor more specifically restricted to the sensitivity of the BIS. The BIS itself is defined in terms of sites of action of anxiolytic drugs. In the case of the novel anxiolytic drugs, acting at 5HT1A receptors, we are probably looking at a serendipitous selectivity for serotonergic terminals located in the BIS. When serotonin itself binds to 5HT1A receptors, it would also be binding to many other receptors, served by collaterals in other parts of the defence system. However, the most commonly used classical anxiolytics, the benzodiazepines, operate by modulating the sensitivity of the GABA-A receptor without binding to it themselves. This would be an ideal site at which a circulating 'anxiety-specific' hormone could act. There is evidence for endogenous compounds that bind to benzodiazepine receptors and which could have such a hormonal action (George *et al.* 1994; Lamacz *et al.* 1996; Sudakov *et al.* 2001; Aufdembrinke 1998; Ozawa *et al.* 1994). We might expect, then, that longer-term changes in the reactivity of such a system could lead to a personality factor that would influence specific morbidity for generalized anxiety disorder and would not affect morbidity for obsessive compulsive disorder, panic disorder or depression.

General modulation

That personality factors should operate at a more global level also makes evolutionary sense. Specific responses (risk assessment, panic) are appropriate to particular levels of threat. But there is a delicate balance in normal ecological situations between general risk-proneness (which could cause you to be killed by a predator before you reproduce) and risk aversion (which could cause you or your offspring to starve to death while avoiding a predator and so, again, fail to reproduce). Specific learning tied to local stimuli will deal with specific special risks, but we would expect that, in addition, there would be longer-term feedback mechanisms that adjust non-specific risk-proneness both within an individual and within a genetic pool. Critically, such long-term adjustments cannot be specific to a particular defensive distance (and hence symptomatology). They must modulate the overarching factor of defensive distance itself.

In principle, there might be a wide range of negative affective events combining to create a general avoidance tendency reflecting activation of a single FFFS. Conversely, positive affective events would activate the BAS. Activation of inputs to FFFS and inputs to BAS will sum to produce general arousal (Gray and Smith 1969) and subtract to produce behavioural output. The more similarly the FFFS and BAS are activated, the more conflict will result and activate the BIS. This will increase arousal further and bias decisions towards avoidance (Gray and McNaughton 2000).

Gray and McNaughton (2000) and McNaughton and Corr (2004) deliberately held back from specifying the relationship of the components of the revised theory and personality factors; we now incline to the view that the old 'Anxiety' axis (i.e., neurotic-introversion) should be relabelled as 'Punishment Sensitivity', or 'Threat Perception', or simply 'Defensive Distance', with lower order factors of this orthogonal 'dimension' breaking down into specific, lower-order, oblique FFFS-fear and BIS-anxiety factors. It should be noted here that we encounter another asymmetry: 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.

Two important issues spring from these conclusions: (a) the (inter) dependence of the Punishment Sensitivity (fear + anxiety) and the BAS; and (b) the relationship between fear and anxiety measures on behavioural and psychophysiological measures in the laboratory (in this latter case, it is important to remember the distinction between (i) the *subtractive* nature of reward and punishment on the 'decision' mechanism, *direction*; and (ii) the *additive* effect of reward and punishment on the arousal component, *intensity*).

FFFS/BIS and BAS interactions

An important point for RST is that the theory focuses on *state* changes and considers three basic scenarios: approach, avoid and conflict. But there is a layer of complexity that is old (Gray and Smith 1969) 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 (that is, when there is little conflict between approach and avoidance), they nonetheless interact: this interaction is symmetrical. Activation of one system inhibits the other with respect to decision-making. This inhibitory interaction (in its purest form counterconditioning of one stimulus by a motivationally opposite stimulus) is insensitive to anxiolytic drugs and so is in practice as well as theoretically independent of the BIS (McNaughton and Gray 1983). 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 *behavioural output*. The primary symmetrical interactions between the systems are also non-linear (see below), accounting for such phenomena as behavioural contrast and peak shift (Gray and Smith 1969). Joint activation increases arousal while producing a subtractive effect on the decision process of the model.

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 the aversive system but not the appetitive one. Under conditions of conflict, then, it increases risk aversion.

Corr (2001, 2002a) argued that much of the human experimental data designed to test the BIS and BAS are consistent with the *Joint Sub-systems Hypothesis* of BIS/BAS effects (for a review of this limited literature, see Corr 2004). Interactions are often found between psychometric measures of the BIS/BAS in predicting behavioural effects. The state account of the theory presented above, however, essentially retains the *Separable Sub-systems Hypothesis* (i.e., BIS/BAS effects are functionally independent) of the 1982 version of the theory. While the systems are *neurally* independent, and can be assessed for separate trait sensitivities, their *outputs* will interact when they are *concurrently activated*. We argue that such concurrent activation is usual, but not necessary, under typical human laboratory conditions (e.g., mixed reward/punishment stimuli, weak stimuli) on tasks sensitive to motivational influences.[2]

Separability and dominance of systems

There are two distinct issues here. The first is the issue of intrinsic separability of the systems, and the second is the idea of dominance. Under normal ecological circumstances, the 2000 theory assumes (as did the 1982 theory) that an approach or an avoidance tendency will often capture response mechanisms. But, where approach and avoidance are too evenly matched for straightforward capture, activation of the BIS will enhance the avoidance tendency and so usually lead to avoidance. This might seem to imply that either the BAS or FFFS will be dominant at all times (with the FFFS needing help from the BIS on occasion). However, there are two scenarios where this dominance will be less than absolute. First, is when activation of the BIS changes a weak net approach tendency into an only marginal avoidance tendency.

[2] They are other ways in which the systems may interact to produce complex forms of behaviour. Corr (2002b) noted that frustative non-reward should be generated first in those individuals sensitive to reward (i.e., those who are highly BAS-sensitive), and that the detection of 'non-reward' (i.e., a lower frequency or magnitude of reward than expected) should serve as an input to the BIS (which generates the aversive state; this position is consistent with the Arousal-Decision model presented by Gray and Smith 1969). The experimental prediction is that such a state should be highest in BAS+/BIS + individuals and lowest in BAS –/BIS – individuals. Despite some initial work (e.g., Carver 2004), this prediction has yet to be adequately investigated. Clarity on this matter, as well as others, may, however, need to wait for adequate psychometric measures of the revised FFFS and BIS.

Under these circumstances the observed behaviour will be dominated by risk assessment and exploration (of the external world or of memory) with prepotent approach and avoidance tendencies both suppressed. Second, is when we view behaviour on a longer timescale. In a straight alley, in which both food and shock have been experienced in the goal box, a rat will initially run towards the goal since approach gradients are shallower than avoidance gradients (Miller 1944). This approach to the goal (since it involves passive avoidance) will engage the BIS and so slow approach even more than would the subtraction of the avoidance tendency from the approach tendency. If the memory of the shock is sufficiently aversive, the rat will stop at some distance from the goal box, turn and move away. From this point, the rat is engaged in active avoidance, the BIS is no longer engaged, and so the memory of the shock is perceived as relatively less aversive than when approaching. The rat therefore reverses its direction. This relatively fast switching between states, coupled with the assumption of behavioural momentum, explains the dithering observed in rats during approach-avoidance conflict in runways and that is experienced cognitively, and not always behaviourally silently, in ourselves when faced with difficult choices.

This analysis would deliver the results predicted by the Joint Subsystems Hypothesis provided that, in the vast majority of human experiments, there is simultaneous weak activation of appetitive and aversive systems. Given the presence of goal gradients, there is little reason to assume that, across the whole task, one system dominates the other. Rather, the FFFS, BIS and BAS may be simultaneously activated and the control of behaviour pass from one to the other as a result of the weakness of activating stimuli, variations in memories currently being recalled or, as is often the case, changing task demands.

Different forms of interaction

To say that the systems are fundamentally independent does not mean that their effects on behaviour will be independent. This, in turn, means that assessment of underlying personality factors will involve variables that are likely not to be factorially pure. When tested at the state level, appetitive and aversive systems will frequently be co-activated – albeit unintentionally. Omission of reward is punishing and so it can be difficult to arrange a truly pure reward schedule. With concurrent activation of the systems (one weaker and the other stronger) the more there is heightened activity in one system the more there will be a general suppression of the other (Gray and Smith 1969). This joint sub-systems view is, at a fundamental neural level, wholly consistent with Gray's original view of the two critical personality factors as reflecting independent *sensitivities* to punishment and reward. Sensitivity of one system can be assessed by a carefully purified test, in the absence of contamination from the other, and the personality factor loadings extracted from such pure tests will be independent. The interaction between the systems when they are co-activated even modestly can, however, be complicated.

Pure activation of the BAS or FFFS involves cognitions that, if they lead to action, will result in pure approach (BAS) or pure avoidance (FFFS) of some situation without any tendency to produce cognitions of the opposite affective valence. In such situations, trait differences in the reactivity of the inactivated system will not affect responses to the activated system.

Unequal activation of the BAS and FFFS involves strong activation of one system with weak activation of the system inducing the opposite tendency. Here trait increases in the reactivity of the less activated system will result in decreased cognitive and hence behavioural output from the more activated system and a contrasting increase in arousal. Trait decreases would have the opposite effects. This interaction in the output of the co-activated systems is symmetrical – it does not matter which is the weak and which the strong system. In human experiments this interaction could appear as an attentional bias. The cognitions of the more weakly activated system would be made even weaker and so less able to capture attention. If one is waiting for the executioner's bullet (FFFS activation) then news of a US$100 wage bonus is unlikely to lead to much (BAS-mediated) pleasure!

Similar activation of the BAS and FFFS produces conflict that has effects over and above the interactions produced with unequal activation. The more equal is the activation of the opposing tendencies, the more we have conflict. At low levels, conflict results (via mild BIS activation) in an amplification of the effects of the FFFS on behavioural output while the BAS is controlling behaviour (but not vice versa). Conflict changes the behavioural output qualitatively when the approach and avoidance tendencies (behaviourally silent or not) are sufficiently balanced to make a decision between approach and avoidance difficult to make on the basis of choosing that one which is clearly the more activated. At this point, both approach and avoidance behaviours (as opposed to the positive and negative cognitions represented by activity in the BAS and FFFS) are blocked and exploration and risk analysis are initiated to gather information, positive or negative, that will resolve the conflict. Exploration and risk analysis are most easily detected as behaviour but it is a crucial feature of the theory that they involve behaviourally silent scanning of memory as well as of the environment. Here, trait changes in any of the systems will alter cognitive and so, often, behavioural output. They will also, as a result of changes in cognition, change memory and so future behaviour. Changes in the BAS will alter the external stimulus values at which a BAS/FFFS balance results in conflict. Changes in the FFFS will alter both this balance and, probably, the effect of output from the BIS since, as the theory stands at present, BIS output will often amplify FFFS activity. Increased sensitivity of the BIS will alter the balance by triggering increased FFFS activity at lower levels of conflict. Indeed, the theory attributes some cases of generalized anxiety disorder to excessive output from the BIS that results in negative cognitive bias, i. e. excessive activity in the FFFS for a given input during approach to a (perhaps very mildly) threatening situation.

Testing the systems

As we have seen, the 2000 theory argues for independence of the systems in the sense that all three systems have their own unique, non-overlapping biological control: (a) anxiolytic drugs affect BIS but not BAS nor FFFS;[3] (b) panicolytic drugs affect FFFS but (probably) not BIS; and (c) addictive drugs affect BAS but not BIS or FFFS. The critical point here is that you can change trait features of one system without affecting the outputs from the other systems and detect these changes selectively *provided* you have pure tests of the other systems.

Testing BAS sensitivity without FFFS or BIS is theoretically simple. All that is needed is a task that determines pure sensitivity to reward with no slightest hint of aversive consequences. In practice, some care must be taken as many net positive stimuli have both appetitive and aversive aspects and (a core aspect of the theory) any omission of an expected appetitive stimulus will result in aversion. Error-free learning paradigms (equivalent to the rat finding food in a straight alley) are probably necessary to achieve this. Testing FFFS sensitivity is the same but with the affective signs reversed.

Testing uncontaminated BIS sensitivity is more difficult because the most usual way of generating conflict pits punishment against reward. As a result, changes in sensitivity of either the BAS or FFFS will shift the balance of the conflict and so alter the apparent output of the BIS in that situation. The same is true with variations in the stimuli used in a task.

[3] This description is of the classes of drugs rather than of individual drugs. Some anxiolytic drugs are panicolytic and addictive but that is as a result of side-effects and is not a necessary feature of anxiolysis since other equally anxiolytic drugs are neither panicolytic nor addictive.

Solution of what is *formally* a passive avoidance task does not require the BIS, provided the nominally competing tendency is weak enough not to generate significant conflict (Okaichi and Okaichi 1994). The simplest task without this problem is two-way active avoidance. This involves an avoidance-avoidance conflict rather than an approach-avoidance conflict. Changes in the sensitivity of the FFFS will therefore affect avoidance, as such, equally for the two locations, and changes in passive avoidance *relative to* active avoidance must be due to changes in sensitivity of the BIS.

The procedure to assess the BIS would be as follows: (a) Test groups of individuals in both a one-way and a two-way active avoidance task at varying levels of shock. Determine the normal variation (if any) in one-way active avoidance learning (or perhaps performance) with shock level – this must be due to variation in activation of FFFS without any change in the 'sensitivity' of the BIS. (b) Then look at the additional differences between individuals in two-way avoidance. The drug data say that low BIS sensitivity will result in 'improved' two-way avoidance learning relative to one-way, since anxiolytic drugs improve the former and do not change the latter. Thus, the drugs allow active avoidance, in the two-way situation, to occur without interference from the normally competing passive avoidance tendency. Elsewhere, we have presented other state challenge tests to assess the reactivity of FFFS and BIS modules (McNaughton and Corr 2004).

A computational model

In an attempt to put a bit more flesh on the bare bones of the concepts outlined above, we have constructed a simple computational model. The parameters of the model derive from the interrelations between the FFFS, BIS and BAS that we have outlined above. These interrelations were always implicit in the original (1982) BIS theory. But the revised theory stresses that simultaneous activation of the FFFS and the BAS activates the BIS. This, and the largely ignored symmetrical interactions of the BAS and FFFS, have significant implications for the interdependencies in functional outputs of these systems.

Model specifications

The model (see Figure 5.1) is based primarily on the symmetric BAS-FFFS interactions of the Gray and Smith arousal-decision model. To their basic model we have added the asymmetric effects on these systems of the BIS. (The model is not intended to be quantitative.) Input units

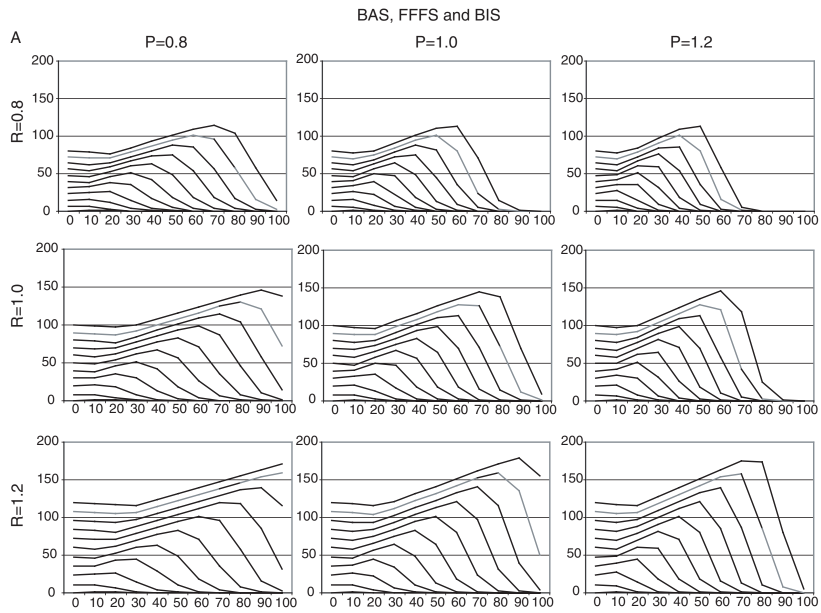

Figure 5.1 Output from the modified Gray and Smith model of Figure 2.6 (see Chapter 2). R, P represent changes in personality factors of reward sensitivity and punishment sensitivity, respectively. The Y axis represents the output of reward-related behaviour. The different lines plotted result from different reward input values (0–100 %). The X axis represents different punishment input values (0–100 %) on the same nominal scale as reward values. A: BAS, FFFS and BIS: output of the full model (see D for block diagram). B: BAS & FFFS (no BIS): output of the model when the BIS component (see D) is eliminated. C: Effect of loss of BIS: difference between the two models in A and B. This represents the type of effect to be expected with people treated with anxiolytics or with very low scores on a factor relating to BIS sensitivity. By contrast, variation due to 'trait anxiety' would be expected to follow the changes in P. D: Modified Gray and Smith model: a block diagram of the model is shown at the D. R_i, P_i are inputs to the BAS and FFFS, respectively. These are multiplied by a sensitivity value ($R = 0.8$, 10., 1.2; $P = 0.8$, 1.0, 1.2 in the graphs) to deliver internal representations R'_i, P'_i. These sum to produce arousal (a). Their difference is input to a decision mechanism that assumes a normal distribution of their inputs with a particular standard deviation (*s*). Conflict detection in the BIS is modelled as a function that increases with increasing activation of the more activated of the two systems (FFFS or BAS) and decreases as the unsigned difference between the two systems increases

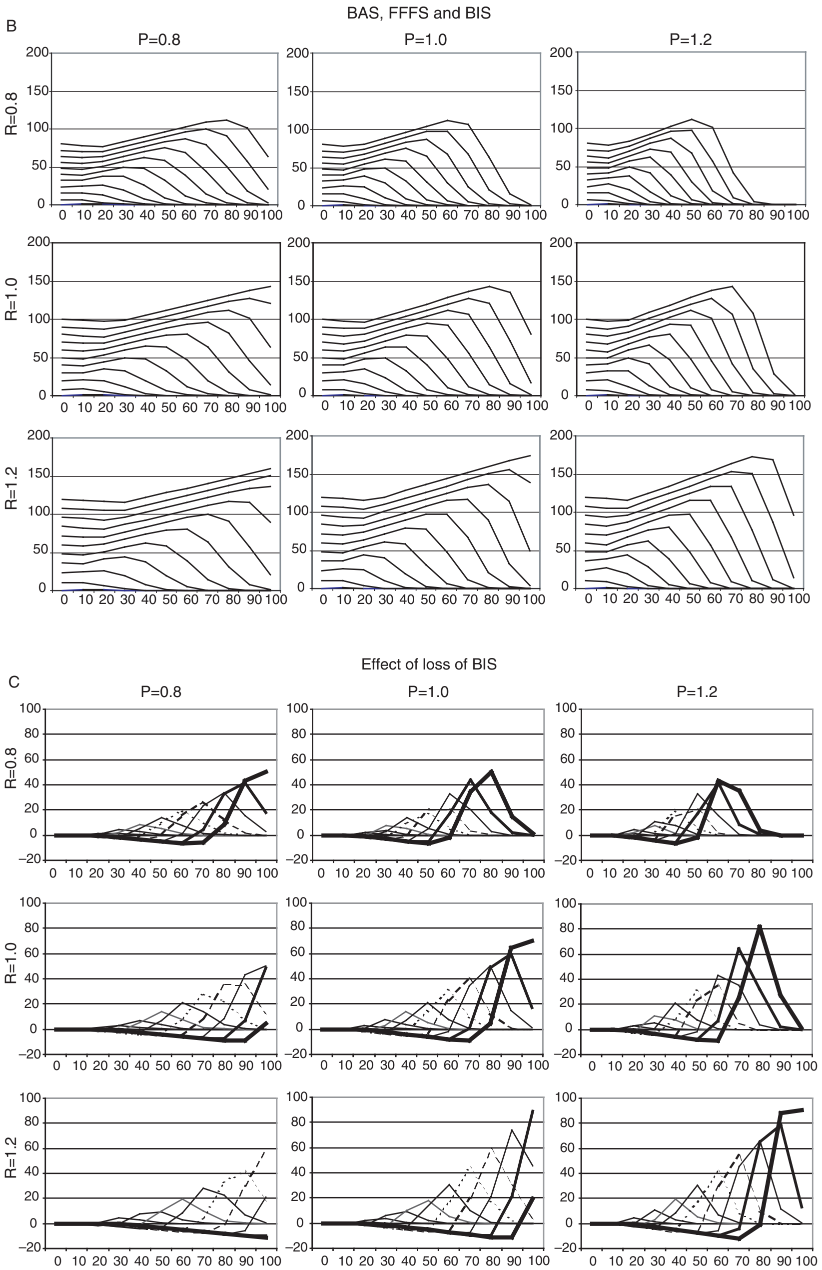

Figure 5.1 (*cont.*)

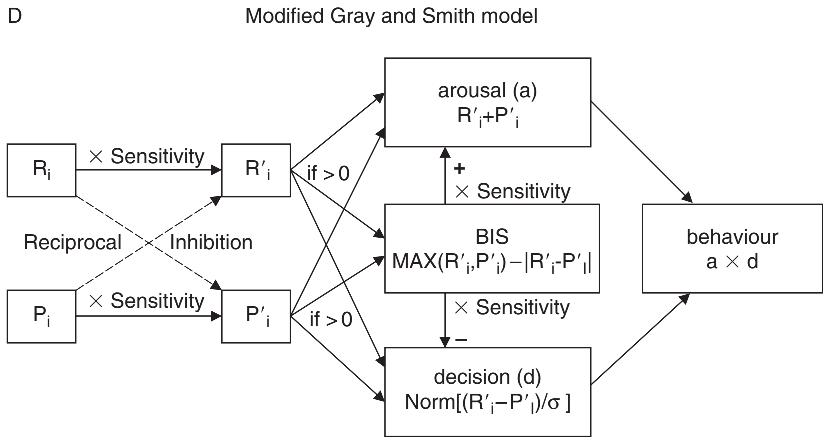

Figure 5.1 (*cont.*)

and response output units are expressed as percentages of some nominal standard. Sensitivities are expressed as proportions of some nominal mean or median of a normal population and are used as multipliers. Thus with reward sensitivity (impulsivity) at the mean for the population (R=1.0) changes in reward input between 0–100 per cent produce changes in output between 0–100 per cent, respectively, provided there is no punishment input. With R=1.2 (representing reward sensitivity 20 per cent above the norm), input between 0–100 per cent produces changes in output between 0–120 per cent, respectively.

Computationally, reward and punishment sensitivities are implemented as multipliers between the external input (R_i, P_i) and its internal representation (R'_i, P'_i). We depart slightly from the Gray and Smith model in that the mutual inhibition of the reward and punishment systems is represented as being fed-forward from the external inputs to the internal representations rather than involving recursive connections. This provides computational simplicity (requiring no recursion) and is likely to be closer to neural reality. The internal representations of reward and punishment (R'_i, P'_i) are then separately summed (to calculate *arousal*) or subtracted (to calculate *decision*). We have made a second minor modification to the Gray and Smith model, here, preventing R'_i or P'_i from taking negative values. This was done to match the fact that neurons cannot be inhibited below a zero firing rate. An important feature, retained from the original model, is that the decision mechanism operates on the assumption of a normal distribution of the strength of its inputs. The arousal and decision components are then multiplied to generate the overall behavioural output observed. This latter step allows generation of such phenomena as behavioural contrast and peak shift when a free operant measure is being used. The separate arousal and decision outputs can be used if, say, autonomic output or, say, choice in a two choice paradigm is to be modelled, respectively.

While the architecture of the model is quite specific, its outputs are not tightly constrained, except in general form. Not only are the input and output values expressed in arbitrary units but the model contains a large number of free parameters. Figure 5.1 shows the effects of explicitly varying reward and punishment sensitivity by 20 per cent up and down from the nominal norm. Other parameters that could be varied are: the extent of reciprocal inhibition between R'_i and P'_i; the sensitivity of the BIS to detect conflict; the extent to which output from the BIS affects arousal and decision, respectively; and the way that arousal and decision combine to deliver predicted output. What the model does therefore is illustrate only the general overall form of the kind of interactions that can be expected between the systems in their generation of output. Any attempt to produce a neurally faithful model should start with separate estimations of individual components (such as reward sensitivity) and only later combine these (now parametrically rigid) components into a full model.

Two features are of particular interest. First, is the non-linear behaviour of the basic model (even without the BIS contribution). This is the result, transferred from the original Gray and Smith model, of including the assumption of a normal distribution of inputs to the decision mechanism. Second, is the prediction that changes in the BIS can produce (at least small) opposite effects at certain parameter values to those that would normally be expected. Such effects have very occasionally been observed by one of us (Monahan (1989)) with anxiolytic drugs in animals.

The output of the model represented in Figure 5.1 is focused on a single free operant behaviour generated by the reward system. Activation of the BAS, uncontaminated by activation of the FFFS, is represented by the outputs graphed when input to the punishment system is zero (i.e., the left-most point of each curve). This is represented as being a simple linear relationship between input and output. In practice this would be expected to be at least a sigmoid function. Any attempt at a quantitative version of the model should, therefore, start with a purely appetitive task and systematically vary reward value in a substantial number of steps that cover the entire normal range of the input.

Output corresponding to a single behaviour generated by the punishment system would be identical in form to that represented for reward, in the absence of a BIS – but with the axes reversed. The effect of the BIS would, of course, be the opposite of that shown in Figure 5.1. Again, any attempt at a quantitative version of the model should start with a purely aversive task and systematically vary punishment value in a substantial number of steps that cover the entire normal range of the input.

The estimation of the free parameters in the model and of the linearity or otherwise of the BIS-arousal and BIS-decision interactions could proceed in a similar fashion once the basic reward and punishment functions had been determined. So while there are a large number of free parameters, each can be independently constrained by data.

Hidden complexity

Consideration of the interdependence of the FFFS and BAS – or the BIS and BAS in the 1982 version of RST – reveals the complexity of prediction. Corr (2004) noted that RST has received mixed support from laboratory studies, which stems in part from a failure to translate theoretical constructs into operational variables amenable to a fair test in the human laboratory. This issue is especially important when we ask the question: what empirical findings would disconfirm modified RST? (See Matthews, chapter 17.) At first sight it may seem that there have been (too) many personality-reinforcement associations that are prohibited by RST, whether conceived in terms of the Separable Sub-systems Hypothesis or the Joint Sub-systems Hypothesis. But sometimes this apparent predictive failure reflects the finer details of the theory being neglected. For example, RST allows the situation in which the presentation of punishment leads to enhanced BAS-related behaviour (and, indeed, was based on animal data that demonstrated this phenomenon, such as behavioural contrast and peak shift). For example, in a study of sexual response, Barr and McConaghy (1974) found that anxiety enhanced appetitive electrodermal conditioning. This experimental outcome is permitted within RST because the presentation of punishment during the performance of a BAS-controlled behaviour has two effects: (a) arousal-induced enhancement of any ongoing response, and (b) inhibition of behaviour. It is a matter of experimental detail which of these effects is witnessed: with relatively weak punishment, the arousal enhancing effect may more than compensate for behavioural inhibition (which, in any event, may not be so apparent in some psychophysiological measures, e.g., eye-blink conditioning). These possible effects are evident from the original Gray and Smith (1969) model.

It is often possible in RST terms to provide a post hoc account, even for seemingly contradictory findings; but this is an unsatisfactory way to test theory. Much better is (a) rigorous specification of parameters of any given task; (b) consideration of mutually inhibitory effects of the FFFS and BAS; (c) operational specification of prevailing reinforcement; and (d) consideration of level and effects of non-specific arousal induction. With this information in hand, it is then possible to *predict* (rather than postdict) the outputs of the FFFS, BIS and BAS in any given experimental situation. Even so-called, simple learning tasks (e.g., eye-blink conditioning) hold hidden complexities. A priori hypothesis formation and rigorous testing are required to avoid the possibility that 'the new RST may perhaps afford excessive "wriggle room" for explaining unexpected effects' (Matthews, chapter 17).

Existing RST personality measures

The small amount of evidence available suggests that genetic and long-term environmentally-induced changes in a combination of noradrenergic and serotonergic systems could underlie a personality factor that is normally measured as 'neurotic-introversion' (*ex hypothesi*, 'Punishment Sensitivity': a shift along the dimension of defensive distance for any fixed level of external threat) – reflecting principally the activity of the FFFS (but, due to the close relationship between the FFFS and BIS, also sharing variance with BIS-related anxiety). At the neural level, this factor would represent significant co-variation of noradrenergic and serotonergic function. Co-variance could result for two reasons that are not mutually exclusive:

(a) if the primary agent were monoamine oxidase or stress this would act jointly on the two monoamines;
(b) synergistic changes in both systems may be required to alter behaviour appropriately.

Conceptualization of the factor as punishment-related rather than as 'trait anxiety' is driven by two facts. First, at the neural level, the factor is related to risk not only for generalized anxiety but for many other pharmacologically separate conditions (see McNaughton and Corr, chapter 2), including depression. An increased sensitivity to higher levels of punishment (spanning both the fear and anxiety systems), then, has much in common with the older concept of neurotic disorders. Secondly, at the personality level, 'trait anxiety' would seem to suggest permanent anxiousness. In fact, it is only a risk factor for the later development of anxiety-related disorder. Here it is worth noting that Taylor himself did not see 'manifest anxiety' (which we can identify with punishment sensitivity/neurotic introversion) as related in any direct way to clinical anxiety and thought 'the test might better have been given a more non-committal label' (Taylor 1956). One issue to be dealt with in relation to 'anxiety' measures, such as the Spielberger Trait Anxiety scale is that items span a range of FFFS and BIS situations, and do not reflect anxiety as conceptualized here. A second is that they share only about 50 per cent of variance with Eysenck's Neuroticism.

What of the scales developed in recent years to measure the FFFS, BIS and BAS? These have been developed to replace more general ones of anxiety and impulsivity that provided convenient names for Gray's rotation of Extraversion and Neuroticism. Some are related to the systems' functioning (e.g., the BIS/BAS scales; Carver and White 1994); others to general expectancies of reward and punishment (e.g., the Generalised Reward and Punishment Sensitivity scales; Ball and Zuckerman 1991); and still others to the characteristic (animal-analogue) behavioural outputs of these systems (e.g., the Gray-Wilson Personality scales; Wilson, Barrett and Gray 1989).

To demonstrate how far the revised theory has gone in clarifying the distinction between the FFFS (fear) and BIS (anxiety), consider the BIS items from the highly popular Carver and White BIS/BAS scales (hypothesized FFFS/BIS designations shown in brackets):

(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)

This 'catch-all' scale has proved highly popular in experimental studies, perhaps because it does mix FFFS and BIS items, and thus measure the more general construct of 'Punishment Sensitivity'. However, the need to distinguish between fear and anxiety is important in the new theory, because, in some situations, FFFS-fear and BIS-anxiety control opposite motivational tendencies (i.e., avoidance vs. cautious approach). Therefore, we are likely to need separate scales that are sensitive (a) to perceptual sensitivity (input variables), comprising defensive distance and general punishment sensitivity (also general appetitive motivations, so corresponding to perceptual distance to reward); and (b) processes (outputs) of the FFFS (e.g., avoidance), BIS (e.g., risk assessment) and BAS (e.g., exploration).

Implications of revised RST for testing

Revised RST holds a number of important implications for the laboratory testing of personality hypotheses (Corr and Perkins 2006; McNaughtion and Corr 2004). For example, consider a standard question in the psychophysiology of emotion: what are the psychophysiological correlates of fear and anxiety? The conventional psychophysiological approach to personality is to take a psychophysiological measures (e.g., EMG startle) and relate this measure to psychometric traits (e.g., trait anxiety). At best, approximate relations may be found, for example, between arousal and the BIS. The problem with this approach is the *atheoretical* nature of the relationship between personality and psychophysiological parameters. As shown by the discussion of 'defensive distance' (see McNaughton and Corr, chapter 2), a threat stimulus of a fixed intensity leads to different behavioural reactions depending on the individual's *perceived* defensive distance; and with each distinct defensive behaviour (e.g., avoidance vs. freezing) different psychophysiological processes are engaged. With psychophysiological measures that may measure whole defensive system functioning (e.g., skin conductance), this may not be too much of a problem. But it is altogether a different matter when we want to measure activation of specific neural modules, or even to distinguish between fear and anxiety. The widely reported 'fractionation' (Lacey 1967) of psychophysiological measures may be a result of the activation of different neural modules at different defensive intensities.

We should also expect, though, that the introduction of reinforcement during the performance of a cognitive task would have definite consequences, but the precise pattern of effects observed would depend on the nature of the reinforcement and cognitive parameters of the task. As already discussed, specific modules have specific reactions, and we would need to know in advance the precise cognitive demands before predicting outcomes. The observation that cognitive parameters interact with personality traits does not strengthen or weaken RST.

An important conclusion of RST is that it should be possible to separate different syndromes of defensive disorder by using theoretically-based challenge tests and so by-pass the problem that (given the interconnectedness of structures) different syndromes can present with much the same symptoms. Indeed, a key feature of the tests we propose is that they should seldom be directed towards the most obvious symptoms and should be administered when state anxiety and hence symptoms are minimal. The same would, of course, be true of any challenges used to activate the brain for imaging (McNaughton and Corr 2004). The central idea behind the suggestion for differential diagnosis is that the specific nodes of the defence system should be selectively challenged to determine whether they are functioning normally. Such challenges should be designed to produce *minimal* reactions from the rest of the defence system. Otherwise, anxiety (or fear or panic) will automatically spill over into activation of much of the remainder of the system, so making it impossible to determine at which point excessive reactions begin. An important corollary of this recursiveness (and an idea gradually creeping into conventional diagnosis) is that co-morbidity is likely to be extensive. For there is little reason to suppose that just one node of the overall defence system should often be the only one overreactive in any one individual at any one time.

Testing specific neural modules in the defensive hierarchy

Considering the FFFS first, starting at the bottom of the defence system with the periaqueductal gray, what we require is a stimulus maximally activating this region accompanied by minimal activation of other parts of the defence system. With such a challenge we could then test patients for the extent to which the periaqueductal gray itself is overreactive, as opposed to being secondarily triggered by excessive activity elsewhere in the defence system. The periaqueductal gray controls 'fight-flight reactions to impending danger, pain, or asphyxia' (see McNaughton and Corr, chapter 2). 'Danger' in any general sense could clearly produce widespread activation of the defence system before activating the periaqueductal gray. To detect not only clinical panic disorder (which some define as involving anxiety), but also those who show panic without anxiety, one could determine the *threshold* level of CO_2 required to elicit an attack. More subtle assessment could be necessary; and, indeed, it seems that panic disorder may be detectable from irregularities in respiratory rhythm and perhaps the response to respiratory challenge. As soon as panic is elicited, other parts of the defence system could contribute to the attack. So, challenge with fixed levels of CO_2 is not only theoretically unattr-active but does not discriminate panic well from, e.g., specific phobias. Threshold measurements, on the other hand, should detect supersensitivity in the periaqueductal gray independent of other abnormalities in the defence system. There may also be relatively input-specific abnormalities of the periaqueductal gray whose detection would require testing with, say, painful stimuli or adrenaline challenge as well as asphyxia.

We have linked amygdalar dysfunction with the arousal component of anxiety. The most obvious relevant challenge would be fear-potentiated

Table 5.1 *A sample of experimental assays to measure the activity in the Fight-Flight-Freeze System (FFFS), Behavioural Approach System (BAS) and Behavioural Inhibition System (BIS)*

Motivational system	Experimental assays
FFFS:	One-way avoidance Anticipatory arousal (e.g., electrodermal activity) Conditioned freezing (no conflict) – electromyographic Cold pressor test Hyponeophagia (inhibition of eating in novel environment) Serotonin challenge
BAS:	Simple approach (e.g., CARROT task)[a] Reaction time to appetitive cue (vs. neutral cue) Attentional bias to appetitive stimuli (e.g., dot probe) Reactions to omission/termination of punishment (e.g., psychomotor activity) Error-free learning Dopamine challenge
BIS:	Approach-avoidance conflict (classic test; e.g., interpersonal interaction, 'performance anxiety') Avoidance-avoidance conflict (pure test; e.g., flight vs. freezing) Approach-approach conflict (frustration test) Counter-conditioning Two-way avoidance (low anxiety = better performance) Q-task (behavioural inhibition)[b] 'Fear' (anxiety)-related startle (arousal potentiation) Geller-Seifter test (frequency of conditioned response with aversive stimulus) Vogel conflict (frequency of conditioned consumption with aversive stimulus) Extinction Reversal learning Benzodiazepine agonist challenge

Note: [a] The CARROT task was developed by Powell, Al-Adawi, Morgan and Greenwood 1996.
[b] The Q-task was developed by Newman, Wallace, Schmitt and Arnett 1997.
References and descriptions of the other essays may be found in Flint (2002, 2004) and Pickering, Corr, Powell, Kumari, Thornton and Gray 1997.

Table 5.2 *Summary of 1982 and 2000 versions of Reinforcement Sensitivity Theory (RST) and suggested personality scales and neural systems*[a]

		1982 theory	2000/2004 theory
	Adequate input	UCS-Pun+, UCS-Rew−	Punishment of all kinds: UCS-Pun+, CS-Pun+, UCS-Rew−, CS-Rew−,
	Output	Avoidance, defensive attack	Avoidance, freezing, defensive attack
FFS/FFFS:	**Neurochemistry**		
	Emotion	Panic and rage	Panic, phobia, rage, fearfulness
	Trait	Psychoticism	Fear ('neurotic-introversion')
	Adequate input	CS-Rew+, CS-Pun−	Reward of all kinds:[b] *CS-Rew+, UCS-Pun−, CS-Pun−*
	Output	Approach, active avoidance	Approach, active avoidance
BAS:	**Neurochemistry**		
	Emotion	Anticipatory pleasure, 'hope'	Anticipatory pleasure, 'hope'
	Trait	Impulsivity (purpose-built BAS scales; see text)	'Impulsivity' (purpose-built BAS scales; see Chapter 1)
	Adequate input	CS-PUN+, IS-Pun+, CS-Rew−, IS-Rew− Novelty (IS-Rew+/IS-Pun+compound)	Conflict stimuli of any kind (e.g., CS-Rew+/ UCS-Pun+)
	Output	Passive avoidance, extinction enhanced information processing, arousal	Passive avoidance, risk assessment, enhanced information processing and arousal
BIS:	**Neurochemistry**		
	Emotion	Anxious rumination of impending danger	Anxious rumination of impending danger
	Trait	Anxiety ('neurotic-introversion')	Anxiety (but not 'neurotic-introversion' per se)

Note: [a] FFF = Fight-Flight System; FFFS = Fight-Flight-Freeze System; BAS = Behavioural Approach System; BIS = Behavioural Inhibition System. UCS-Pun+ = unconditioned (innate) fear stimulus; UCS-Rew − = unconditioned omission/termination of expected reward; CS-Rew+ = conditioned appetitive stimulus; UCS-Pun − = unconditioned 'relief of non-punishment'; CS-Pun − = conditioned 'relief of non-punishment' stimulus; CS-Pun+ = conditioned fear stimuli; IS = Pun+ = innate anxiety stimulus (e.g., cat odour); CS-Rew − = conditioned frustrative stimuli; IS-Rew − = innate frustrative stimuli.

[b] The BAS is not formally involved in controlling reactions to unconditioned appetitive stimuli (i.e., UCS-Rew+).

startle, since this is not only sensitive to anxiolytic drugs (including when injected into the amygdala),[4] but is also insensitive to hippocampal lesions.

Next we come to the septo-hippocampal system. What is required is a test sensitive to septo-hippocampal system damage and anti-anxiety drugs, but *not* to amygdalar or periaqueductal gray lesions. The most obvious tasks, here, are spatial navigation, delayed matching to sample and behaviour on a fixed interval schedule of reward. Of these, delayed matching to sample can be most clearly set up in an anxiety-free form and so would probably be preferable, but it might be too specific in the aspects of septo-hippocampal function which it engages.

Global activity of FFFS, BAS and BIS

A number of promising laboratory tasks and naturalistic procedures have either already been designed or are suggested by careful consideration of the details of the theory to provide empirical tests of RST. These tests are important in verifying predictions of the theory but, equally important, they are required to provide opportunities for disconfirmation which is necessary for the theory to develop – this is important in RST which can be seen to be in a state of continuous development. Table 5.1 shows some of the existing test and possible tests that may be used to index the sensitivity of the three systems.

Out of the woods: putting it all together

Description of the neuropsychology of the FFFS, BAS and BIS is complicated. In an attempt to provide some degree of clarity, Table 5.2 show the differences between the old and new RST versions, as suggested by the above analysis.

Conclusion

Assuming that RST has correctly specified the neuroanatomical bases of defensive and approach systems, it has yet to specify how these systems relate to overt behaviour and individual differences in overt behaviour, namely personality. What was called 'Anxiety' (i.e., running from $E-/N+$ to $E+/N-$) can no longer be so labelled. This axis may either be (a) FFFS-fear,

[4] One of Gray's PhD students, Jasper Thornton (1998), reported that, in healthy volunteer subjects, 'fear-potentiated' startle was selectively reduced by the anxiolytic drug, diazepam (15 mg oral) – this effect was not, however, observed with 5mg (also see Patrick, Berthot and Moore, 1996).

or (b) Punishment Sensitivity or Defensive Distance or Threat Perception (incorporating FFFS and BIS). However, given the independence of the FFFS and BIS, we should expect significant and important individual differences in BIS sensitivity and functioning that are independent of FFFS sensitivity and functioning. In addition to the obvious links between FFFS and fear/phobia, and BIS and generalized anxiety, we could imagine other varieties of individuals, for example, an individual with a weak FFFS and BAS who had a hypersensitive BIS, who would ruminate in a non-emotional way about almost anything; or a hyperactive FFFS individual, highly prone to fear, but where conflict does not activate the hypoactive BIS, leading to pure non-ruminative fear; or a BIS-insensitive, BAS-sensitive individual with a weak FFFS who may be especially prone to psychopathic-type behaviour. Arguably, the distinction between FFFS-fear and BIS-anxiety renders the variety of clinical conditions more amenable to theoretical analysis and explanation.

References

Abel, E.L., Altman, H.J. and Commissaris, R.L. (1992), Maudsley reactive and nonreactive rats in the forced swim test: comparison in fresh water and soiled water, *Physiology and Behavior*, 52, 1117–1119

Aufdembrinke, B. (1998), Abecarnil, a new beta-carboline, in the treatment of anxiety disorders, *British Journal of Psychiatry*, 173, 55–63

Ball, S.A. and Zuckerman, M. (1991), Sensation seeking, Eysenck's personality dimensions and reinforcement sensitivity in concept formation, *Personality and Individual Differences*, 11, 343–353

Barr, R.F. and McConaghy, N. (1974), Anxiety in relation to conditioning, *Behaviour Theraphy*, 5, 193–202

Blizard, D.A. (1981), The Maudsley reactive and nonreactive strains: a North American perspective, *Behavior Genetics*, 11, 469–489

Broadhurst, P.L. (1960), Applications of biometrical genetics to the inheritance of behaviour in H.J. Eysenck (ed.), *Experiments in Personality*, vol. 1, *Psychogenetics and Psychopharmacology* (London: Routledge Kegan Paul), pp. 1–102

(1975), The Maudsley reactive and nonreactive strains of rats: a survey, *Behavior Genetics*, 5, 299–319

Candland, D.K. and Nagy, Z.M. (1969), The open field: some comparative data, *Proceedings of the National Academy of Sciences*, 159, 831–851

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, 319–333

Commissaris, R.L., Verbanac, J.S., Markovska, V.L., Altman, H.J. and Hill, T.J. (1996), Anxiety-like and depression-like behavior in Maudsley reactive (MR) and non-reactive (MNRA) rats, *Progress in Neuropsychopharmacology and Biological Psychiatry*, 20, 491–501

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

(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.

Corr, P.J. and McNaughton, N. (in preparation), Implications and extensions of J.A. Gray's behavioural inhibition system: a prolegomenon to the neuroscience of personality

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

Deakin, J.F.W. and Graeff, F.G. (1991), 5-HT and mechanisms of defence, *Journal of Psychopharmacology*, 5, 305–315

Eysenck, H.J. (1944), Types of personality: a factorial study of seven hundred neurotics, *Journal of Mental Science*, 90, 851–861

(1947), *Dimensions of Personality* (London: Kegan Paul)

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

Eysenck, H.J. (1967), *The Biological Basis of Personality* (IL: Thomas, Springfield)

Eysenck, H.J. and Eysenck, S.B.G. (1975), *Manual of the Eysenck Personality Questionnaire (Adults)* (London: Hodder and Stoughton)

Flint, J. (2002), Animal models of personality in R.P. Ebstein, R.H. Belmaker and J. Benjamin, *The Molecular Genetics of the Human Personality* (Washington DC: APA), pp. 63–90

(2004), The genetics of neuroticism, *Neurosciencce and Biobehavioral Reviews*, 28, 307–316

George M.S., Guidotti, A., Rubinow, D., Pan, B., Mikalauskas, K. and Post, R.M. (1994), CSF neuroactive steroids in affective disorders: pregnenolone, progesterone, and DBI, *Biological Psychiatry*, 35, 775–780

Gray, J.A. (1970), The psychophysiological basis of introversion-extraversion, *Behaiour 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

(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)

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

Gray, J.A. and N. McNaughton (2000), *The Neuropsychology of Anxiety: An Enquiry into the Functions of the Septo-Hippocampal System* 2nd edn, (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

Graeff, F.G. (1994), Neuroanatomy and neurotransmitter regulation of defensive behaviors and related emotions in mammals, *Brazilian Journal of Medical and Biological Research*, 27, 811–829

Flint, J. (2002), Animal models of personality in J. Benjamin, R.P. Ebstein and R.H. Belmaker (eds), *Molecular Genetics and the Human Personality* (London: American Psychiatric Publishing), (pp. 63–90)

(2004), The genetics of neuroticism, *Neuroscience and Biobehavioral Reviews*, 28, 307–316

Hall, C.S. (1951), The genetics of behavior in S.S. Stevens (ed.), *Handbook of Experimental Psychology* (New York: Wiley), pp. 304–329

Hunt, H.F. and Otis, L.S. (1953), Conditioned and unconditioned emotional defecation in the rat, *Journal of Comparative and Physiological Psychology*, 46, 378–382

Lacey, J.I. (1967), Somatic response patterning and stress: some revisions of activation theory in M.H. Appley and R. Trumbull (eds), *Psychological Stress* (New York: Appleton Century Crofts), (pp. 14–43)

Lamacz, M., Tonon, M.C., Smih-Rouet, F., Patte, C., Gasque, P., Fontaine, M. and Vaudry, H. (1996), The endogenous benzodiazepine receptor ligand ODN increases cytosolic calcium in cultured rat astrocytes, *Molecular Brain Research*, 37, 290–296

Lykken, D.T. (1971), Multiple factor analysis and personality research, *Journal of Experimental Research in Personality*, 5, 161–170

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

McNaughton, N. and Gray, J.A. (1983), Pavlovian counterconditioning is unchanged by chlordiazepoxide or by septal lesions, *Quarterly Journal of Experimental Psychology*, 35, 221–233

Miller, N.E. (1944), Experimental studies of conflict in J.M. Hunt (ed.), *Personality and the Behavioural Disorders* (New York: Ronald)

Monahan, A.M. (1989), The involvement of endogenous opiate systems in the anxiolytic actions of the benzodiazepines and melatonin. Unpublished MSc thesis, University of Otago, Dunedin, New Zealand

Newman, J.P., Wallace, J.F., Schmitt, W.A. and Arnett, P.A. (1997), Behavioral inhibition system functioning in anxious, impulsive and psychopathic individuals, *Personality and Individual Differences*, 23, 583–592

Okaichi, Y. and Okaichi, H. (1994), Effects of fimbria-fornix lesions on avoidance tasks with temporal elements in rats, *Physiology and Behavior*, 56, 759–765

Ozawa, M., Nakada, Y., Sugimachi, K., Yabuuchi, F., Akai, T., Mizuta, E., Kuno, S. and Yamaguchi, M. (1994), Pharmacological characterization of the novel anxiolytic carboline abecarnil in rodents and primates, *Japanese Journal of Pharmacology*, 64, 179–187

Patrick, C.J., Berthot, B.D. and Moore, J.D. (1996), Diazepam blocks fear-potentiated startle in humans, *Journal of Abnormal Psychology*, 105, 89–96

Perkins, A.M., Revelle, W. and Corr, P.J. (Unpublished), A reanalysis of H.J. Eysenck's (1994) medical checklist data

Pickering, A.D., Corr, P.J., Powell, J.H., Kumari, V., Thornton, J.C. and Gray, J.A. (1997), Individual differences in reactions to reinforcing stimuli are neither black nor white: to what extent are they Gray? in H. Nyborg (ed.), *The Scientific Study of Human Nature: Tribute to Hans J. Eysenck at Eighty* (London: Elsevier), pp. 36–67

Powell, J.H., Al-Adawi, S., Morgan, J. and Greenwood, R.J. (1996), Motivational deficits after brain injury: effects of bromocriptine in 11 patients, *Journal of Neurology, Neurosurgery, and Psychiatry*, 60, 416–421

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, 320–335

Stouffer S.A., Lumsdaine A.A., Lumsdaine M.H., Williams R.M., Smith M.B., Janis I.L. 1949, Studies in social psychology in world war II: vol 2. The american soldier: combat and its aftermath. Princeton, cited by Broadhurst (1960), page 37

Sudakov, S.K., Medvedeva, O.F., Rusakova, I.V., Terebilina, N.N. and Goldberg, S.R. (2001), Differences in genetic predisposition to high anxiety in two inbred rat strains: role of substance P, diazepam binding inhibitor fragment and neuropeptide Y, *Psychopharmacology*, 154, 327–335

Taylor, J. (1956), Drive theory and manifest anxiety, *Psychological Bulletin*, 53, 303–320

Thornton, J.C. (1998), *The Behavioural Inhibition System and Anxiety in Human Subjects* (unpublished Doctoral Thesis, Institute of Psychiatry, University of London)

Viglinskaya, I.V., Overstreet, D.H., Kashevskaya, O.P., Badishtov, B.A., Kampov-Polevoy, A.B., Seredenin, S.B. *et al.* (1995), To drink or not to drink: Tests of anxiety and immobility in alcohol-preferring and alcohol-nonpreferring rat strains, *Physiology and Behavior*, 57, 987–991

Wilson, G.D., Barrett, P.T. and Gray, J.A. (1989), Human reactions to reward and punishment: a questionnaire examination of Gray's personality theory, *British Journal of Psychology*, 80, 509–515