1 Reinforcement Sensitivity Theory (RST): introduction

Philip J. Corr

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

Past and present

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

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

Foundations of RST

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

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

General theory of personality

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

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

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

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

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

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

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

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

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

Standard (1982) RST

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

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

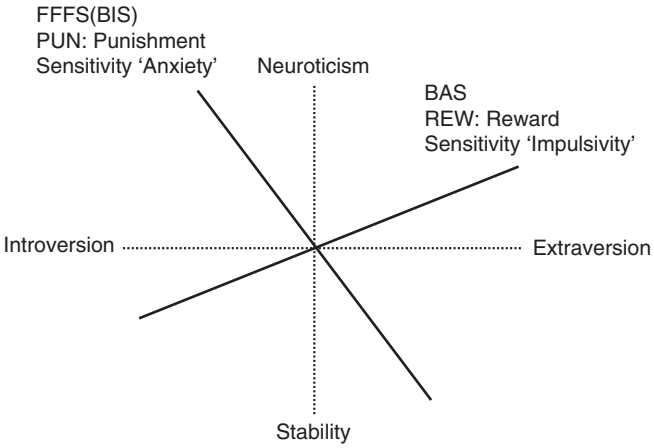

Figure 1.1 Position in factor space of the fundamental punishment sensitivity and reward sensitivity (unbroken lines) and the emergent surface expressions of these sensitivities, i.e., Extraversion (E) and Neuroticism (N) (broken lines). In the revised theory, a clear distinction exists between fear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

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

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

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

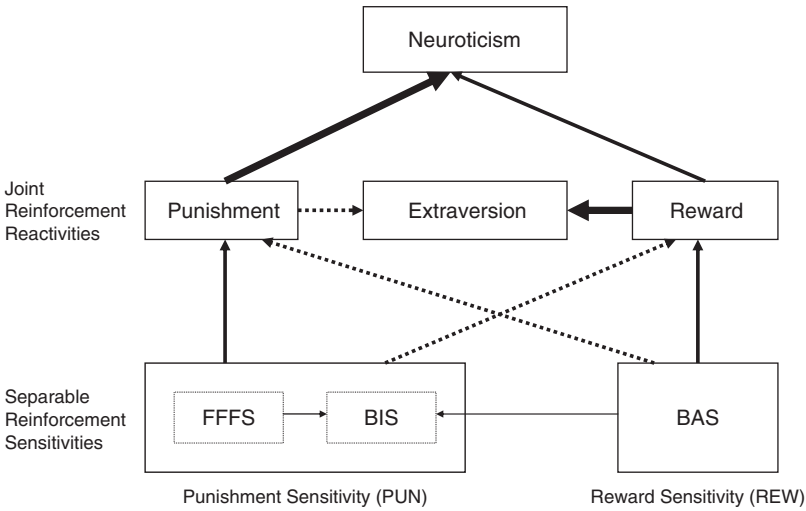

Figure 1.2 A schematic representation of the hypothesized relationship between (a) FFFS/BIS (punishment sensitivity; PUN) and BAS (reward sensitivity; REW); (b) their joint effects on reactions to punishment and reward; and (c) their relations to Extraversion (E) and Neuroticism (N). E is shown as the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

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

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

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

Clinical neurosis

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

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

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

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

Three systems of standard RST

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

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

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

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

Revised (2000) RST

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

Revised RST postulates three systems.

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

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