Neuroscience and Approach/Avoidance Personality Traits:

A Two Stage (Valuation-Motivation) Approach

Philip J. Corr[1] and Neil McNaughton[2]

[1]School of Psychology, and Centre for Behavioural and Experimental Social Sciences (CBESS), University of East Anglia, UK

[2]Dept. Psychology and Neuroscience Research Centre, University of Otago, Dunedin, New Zealand

Corresponding Author

Professor Neil McNaughton,

Department of Psychology,

University of Otago,

PO BOX 56, Dunedin, New Zealand

Telephone: +64-3-479-7643

Fax: +64-3-479-8335

Email: nmcn@psy.otago.ac.nz

ABSTRACT

Many personality theories link specific traits to the sensitivities of the neural systems that control approach and avoidance. But there is no consensus on the nature of these systems. Here we combine recent advances in economics and neuroscience to provide a more solid foundation for a neuroscience of approach/avoidance personality. We propose a two-stage integration of valuation (loss/gain) sensitivities with motivational (approach/avoidance/conflict) sensitivities. Our key conclusions are: (1) that *valuation* of appetitive and aversive events (e.g. gain and loss as studied by behavioural economists) is an independent perceptual input stage – with the economic phenomenon of loss aversion resulting from greater negative valuation sensitivity compared to positive valuation sensitivity; (2) that valuation of an appetitive stimulus then interacts with a contingency of presentation or omission to generate a motivational 'attractor' or 'repulsor', respectively (vice versa for an aversive stimulus); (3) the resultant behavioural tendencies to approach or avoid have distinct sensitivities to those of the valuation systems; (4) while attractors and repulsors can reinforce new responses they also, more usually, *elicit* innate or previously conditioned responses and so the perception/valuation-motivation/action complex is best characterised as acting as a 'reinforc*er*' not a 'reinforce*ment'*; and (5) approach-avoidance conflict must be viewed as activating a third motivation system that is distinct from the basic approach and avoidance systems. We provide examples of methods of assessing each of the constructs within approach-avoidance theories and of linking these constructs to personality measures. We sketch a preliminary five-element reinfor*cer* sensitivity theory (RST-5) as a first step in the integration of existing specific approach-avoidance theories into a coherent neuroscience of personality.

Keywords: personality; economics; anxiety; fear; aversion; conflict; behavioural inhibition; behavioural approach system; fight-flight-freeze system; reinforcement sensitivity theory; loss; risk

Table of Contents

1 INTRODUCTION

1.1 The Goal

With the upsurge of neuroscience in psychology, there has been a proliferation of theories that incorporate personality traits with neural systems that control basic approach and avoidance behaviours. In some cases, this is purely in terms of approach and avoidance (Gray, 1970); in others, approach and avoidance are part of larger schemes (e.g., Cloninger, 1986; Cloninger et al., 1993; for an overview, see DeYoung and Gray, 2009). In fact, the number of these theories has increased rapidly, and members of this extended family include: Depue (Depue and Collins, 1999; Zald and Depue, 2001); Davidson (Davidson et al., 1990; Davidson et al., 2004); and Carver (Carver, 2004; Carver, 2008; Carver and Harmon-Jones, 2009; Carver et al., 2008; Carver and White, 1994). However, as these theories have proliferated, they have tended to become separated from the increasingly complex neural bedrock on which they are nominally based. We recognise this problem as significant for maintaining consensually agreeable definitions of basic concepts, behaviours and underlying systems. In the absence of agreement on these basic issues, it is difficult to know whether differences between theories are substantive rather than differences of definition and/or emphasis.

In this article, we summarise key aspects of what is currently known about the basic state control of approach and avoidance, and the conflict that can occur between them. These are important for theories of personality; and we provide a preliminary translation of the knowledge of these state systems into the realm of personality description and explanation.

1.2 The Problem

Approach-avoidance personality theories invoke long term sensitivities of the major state systems that are activated by appetitive and aversive stimuli, and so attempt to explain consistent patterns of individual differences in behavior. Current theories are not strongly linked to their *a priori* theoretical and empirical foundations. In particular, questionnaires are often constructed intuitively and not validated against more objective neural or behavioural criteria. To tackle this major problem, we argue for the necessity to build a consensus as to the scientific foundations of all approach-avoidance personality theories. It is these general state systems, their interactions, and how they differ between individuals, that provide the facts that are the progenitors of all members of the family of approach-avoidance theories.

The fundamental problem we address is that, in statistical terms, independent trait level variables are the result of interacting state systems. Traits can be viewed as constants within psychological input-output equations. But states, and particularly the behavioural and other measures we use to assess changes in them, are the result of the combination (and often interaction) of the effects of multiple, rapidly changing, variables within an individual.

At the state level, the main problem is theory specification. To test a neuroscientific personality theory, one must take into account the details of the state theory 'equations' through which the trait 'constant' expresses its effects. As we will see (Section 2), this requires careful definition of state level constructs and of their detailed interaction with experimental variables. This issue is complicated by the fact that neural state theories continue to evolve and so their mapping to specific trait measures also needs to evolve.

1.3 The Solution

One solution to this problem is to provide a neuroscientific groundwork that is driven by recent advances in the Reinforcement Sensitivity Theory (RST) of personality (Gray, 1970, 1973, 1981, 1982; Gray and McNaughton, 2000; McNaughton and Corr, 2004, 2008b), which has a lineage dating back to the origins of the current family of approach-avoidance theories of personality. We believe that, while the specifics of each member theory of the approach-avoidance family may currently differ, the fundamental underlying constructs to which they are intended to apply should not – or, if they do, then these differences should be made clear. However, we also believe that the precise nature of these constructs, and of the state interactions between them, remains to be demonstrated experimentally via hypothesis testing of theories such as the preliminary one presented here.

In this article, we will end with an attempt to produce a theory, a revised RST, to indicate possible steps in the direction of integration, so that falsification of it can drive future development. But, our main aims are to provide: (1) a clearer definitional picture of background state concepts, many of them thought to be well-established, that underlie any approach/avoidance-related trait theory; (2) a linkage between these concepts and those of behavioural economics; and (3) a clear (and potentially mathematical) picture of the generation of output from the states that result from the interaction of traits with situational input.

We see the road to progress as starting with the original behaviourist and neural methodologies on which state theory is based and via which it has evolved. We argue that, in humans, travel along this road to a coherent theory of personality will be eased by including methods and theory from the study of valuation as revealed by behavioural economics and extended into the neuroscience realm by 'neuroeconomics' (Glimcher et

al., 2005; Glimcher and Rustichini, 2004; Loewenstein et al., 2008; Sanfey et al., 2006; Zak, 2004). The potential afforded by the union of economics and personality psychology has already been highlighted in several publications (Borghans et al., 2008; Ferguson et al., 2011; Frey and Stutzer, 2007).

The key feature of this approach is that, as has been urged on other grounds, it "involves distinguishing affective value from the requirement for action. That is, it is important to orthogonalize Go, No Go, punishment, and reward, and also the orientation of the action with respect to the cues (to manipulate other aspects of the Pavlovian status of the action), along with the factor controlling whether rewards are related to punishment (eg, money gain vs money loss) or not (eg, money gain vs electric shocks)" (Boureau and Dayan, 2010, p. 16). But, most importantly, we suggest that the terms 'reward' and 'punishment', that have been so prominent in this literature, are used ambiguously – as in "whether rewards are related to punishment". This ambiguity results from the conflation by these terms of independent valuation and motivation stages of processing. We address this issue by providing a preliminary integration of state approach/avoidance theory with fundamental principles and concepts from behavioural economics, which have hitherto been largely absent from this research field.

2 BASIC ISSUES

2.1 Core elements for approach-avoidance theories

There are a number of specific issues, which can be treated independently of specific personality theories, and of each other, that provide the bedrock on which all approach-avoidance theories (state and trait) must build. Some may seem more pertinent

to some theories than others – but, in practice, all theories must take into account the data and methods of analysis that drive the usage of certain 'approach/avoidance' concepts.

The traits of interest to approach-avoidance personality theories have traditionally been linked to 'reward' and 'punishment', which are usually: (a) seen as fundamentally involved in learning; and (b) linked to approach and avoidance, respectively (Gray, 1975). There are two problems here. In relation to their linkage to learning, there is the problem that, a 'reinforcer' produces characteristic innate responses as well as supporting learning and so the capacity for reinforcement is better seen as just one of the properties of a reinforcer (Section 2.2). In relation to their linkage to approach/avoidance, there is the problem that the words 'reward' and 'punishment' are ambiguous in relation to the omission of expected events. Variation in the effects of manipulation of a reinforcer on behaviour can depend not only on the different valuations of gains and losses (Section 2.3) but also on whether the manipulation is presentation or omission and so generates attraction or repulsion (Section 2.4). This raises the issue of how perception/valuation sensitivity interacts with motivation/action sensitivity to control observed behaviour (Section 2.5). Preliminary evidence for separate valuation and motivation sensitivities is presented in Section 2.6 and 2.7.

A practical complication for the assessment of trait attraction sensitivity and trait repulsion sensitivity is that, at the state level, they not only subtract from each other but also have different goal-gradients (Miller, 1944) (Section 2.8). A further complication is that, in addition to these subtractive effects, the inhibition of approach by (approach-avoidance) conflict is neurally distinct from pure avoidance (Gray, 1977). Moreover, the processing of conflict (Section 2.9), and the resultant 'behavioural inhibition', does not encompass all cases where, descriptively, behaviour is inhibited (Gray and McNaughton,

2000). Critically, then, when using terms, such as ‘reinforcement’ and ‘behavioural inhibition’, there needs to be awareness of the various distinct meanings of these terms and of their various implications.

In this article, we distinguish: (a) processing of perceptual inputs (i.e, the valuation of reinforcers that precedes the production of *both* unlearned *and* learned behaviour); (b) the generation of motivated outputs (i.e. the actions that result from attraction and repulsion); and (c) distinct effects on motivated output of the conflict between attraction and repulsion. These distinctions provide us with *five* separate *potential* sources of personality sensitivity that may impact on approach and avoidance behaviour. Although future work may show that these five sources collapse at the level of personality description and causation, we believe that it is prudent to keep them separate so as to test, and thereby potentially refute, their possible separable effects.

The inclusion of valuation from behavioural economics may also sensitize us to the issue that, when testing for effects of these 5 systems, we must also be careful to balance factors such as ambiguity, uncertainty and risk that behavioural economics has demonstrated generate specific aversions, the neural bases of which are already being studied (e.g. Rushworth and Behrens, 2008) and which may have trait components (Sallet and Rushworth, 2009).

A major implication of our analysis is that, because of their valuation/motivation ambiguity, the terms ‘reward’ and ‘punishment’ should only be used with great care and then only when their operational application is make very clear; otherwise they are likely to add to theoretical and operational confusion. As discussed below, we prefer the more descriptive terms ‘attractor’ and ‘repulsor’ for motivating objects (a) to separate them from gain/loss valuation, and (b) to denote their motivational-output functions.

Despite our new perspective, ‘reward’ and ‘punishment’ can be taken to retain their usual meaning *provided* they are concrete events that are presented rather than omitted. In this limited case, they can be treated as positive and negative valuations, respectively, that also lead to approach and avoidance, respectively. The usual theoretical analysis of reinforcement effects holds in this concrete positive case – except that it combines two trait sensitivities for each of reward (gain + approach) and punishment (loss + avoidance). However, with the *omission* of expected events, the terms are best avoided. ‘Punishment’, here, is particularly contentious. Depending on its meaning, it can imply avoidance of danger, approach to safety, or inhibition of a pre-potent responses allowing cautious approach to danger. These three meanings are neurally distinct from each other.

We believe that this change in nomenclature will result in theoretical clarity and that this will more than outweigh the discomfiture of abandoning the familiarity of ‘reward’ and ‘punishment’.

2.2 Reinforcement versus reinforcer

In this section, we tackle the distinction between reinforce*ment* and reinfor*cer* which will come to play a major role in our elaboration of approach, avoidance, and conflict systems underlying personality traits.

Positive and negative reinforcers, as objectively defined concrete stimuli, appear relatively straightforward. For the behaviourist, positive reinforcers increase, and negative decrease, the frequency of behaviours on which they are contingent – with ‘negative reinforcement’ involving the increase in frequency of behaviours by *removal or omission* of the negative reinforcer (Mackintosh, 1974; Millenson and Leslie, 1979).

These definitions spring from animal learning studies in the laboratory. But, while learning theory is the most analytically tractable source of data, it is not the only guide. For the ethologist, organisms approach unconditioned 'positive reinforcers' and avoid unconditioned 'negative reinforcers' (Hinde, 1982). It is at this point that 'reinforcer' becomes a more contentious term. It is usually taken to imply reinforcement, but both types of reinforcer – as concrete objects – can elicit distinctive unconditioned patterns of approach or avoidance behaviour in the total absence of any conditioning history. Innate stimuli can also be used to produce approach-avoidance conflict; and, when they do, they show similar neural (Blanchard and Blanchard, 1972) and pharmacological (Blanchard et al., 1997) sensitivities to those produced with learned stimuli. For our present purposes, therefore, the terms attractor and repulsor seem more appropriate.

2.3 Perception/Valuation – a contribution from behavioural economics

In this section, the case is made for considering valuation as a distinct stage from motivation-output. A reinforcer is an *external motivationally significant event.* One must then confront the issue of individual differences in valuation (both specific exchange rates and more general loss/gain differences). Valuation can often be ignored in the state animal literature since a fixed (at least average) valuation is often taken as a given and not explored within the paradigms used. Animals are food deprived, or other arrangements made, and in most cases the motivational conditions are then held constant and any variation from time to time, or from animal to animal, contributes to residual error. However, it is different in the study of personality, particularly in humans, and any systematic variation in valuation will obviously affect the observed approach or avoidance behaviour.

----- Figure 1 here -----

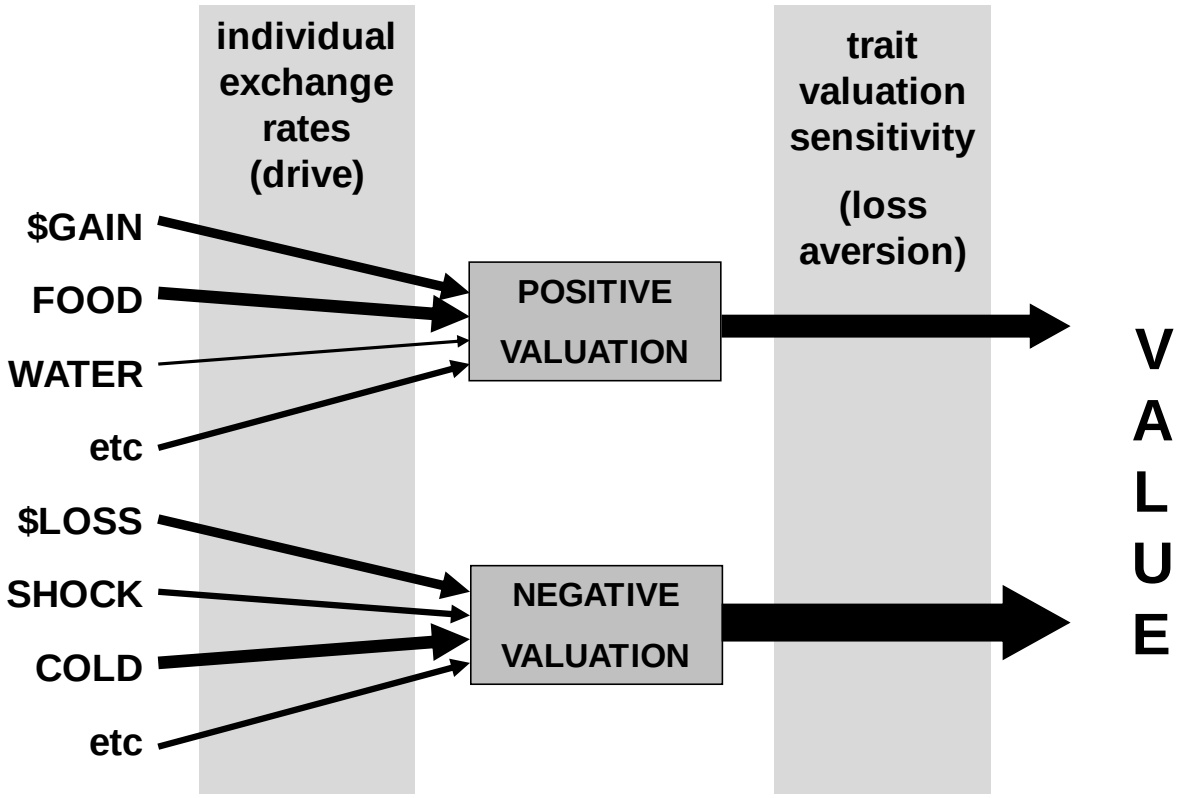

Figure 1. The relationship of external items to their internal value, which controls the strength (but not direction) of their effect on behaviour. Individual items (consumable food, dollar gains, etc) have a value that depends on their amount and the current level of specific drive for that class of item. The amount, therefore, interacts with an individual exchange rate (first grey rectangle, represented by varying arrow thicknesses) to generate an internal valuation, which can be positive or negative depending on the valence of the item (e.g. dollar gain versus dollar loss in the form of removal from an existing store). For the same amount of the same class of item, such as dollars (which necessarily matches individual item exchange rate), negative valence has a higher exchange rate (note thicker arrow) and so generates a greater internal value than positive (second grey rectangle, e.g. loss aversion). Loss aversion is a relative term (comparing the effect of loss with that of the same external value of gain) and we take it to represent the difference between trait gain sensitivity and trait loss sensitivity, with the latter being the greater.

This issue of valuation has already been studied, with a particularly convenient reinforcer, in behavioural economics and, more recently, neuroeconomics (Glimcher and Rustichini, 2004; Loewenstein et al., 2008; Sanfey et al., 2006; Zak, 2004): money. With dollars as the reinforcer, and with the use of a choice paradigm, the external value of a loss (in the form of removal of an amount, such as $1 from an existing store) can be objectively equated with the value of a gain (in the form of addition of the same $1 amount to the store). Responding can be matched in the two cases by comparing, for example, omission of loss to presentation of gain. This has allowed economists to study the internal valuation of losses and gains, the results of which show significant differences between gain valuation and loss valuation (Figure 1).

Of most influence in this field is the work of Kahneman and Tversky (1979) on *prospect theory* – a refined form, 'cumulative prospect theory', was proposed by Tversky and Kahneman (1992). Prospect theory accounts for how people make decisions in situations where they have to decide between alternatives that involve risk (i.e. with uncertain outcomes, but where the probabilities of different outcomes are known). Such studies show that people do not behave according to expected utility theory and their decisions deviate from the strict criterion of classical rationality. The importance for personality theories of this line of work is in its description of how people evaluate potential losses and gains. The main point is that people tend to think in terms of a reference point rather than the final outcome (e.g. total wealth), a phenomenon called 'framing' (e.g. do you prefer to gain a £10 discount or to avoid a £10 surcharge). There is now work that relates elements of prospect theory to neural structure (Trepel et al., 2005) – a paradigm example of the attempt by neuroeconomics to integrate psychology, economics and neuroscience.

Prospect theory describes two processes: editing and evaluation. In the, first, editing phase, possible outcomes are ordered according to some heuristic (people decide which outcomes they see as identical and they then set a reference point and consider lower outcomes as losses and larger ones as gains). In the, second, evaluation phase, people behave as if they can compute a value (or utility) based on potential outcomes and their respective probabilities, and they then choose the alternative having the higher utility.

The main finding of interest to us within prospect theory is that, given the same variation in absolute value (e.g. dollars lost or gained), losses have a larger impact than gains. That is, people on average, and when faced with risk, have a different *sensitivity* to gain (i.e. outcomes above their reference point) compared to loss (i.e. outcomes below their reference point). Studies typically find a loss sensitivity coefficient of approximately two-to-one: people accept risks only if the potential gain is at least twice as much as the potential loss (e.g. Novemsky and Kahneman, 2005). Lower coefficients have also been reported but the basic phenomenon is robust (Sokol-Hessner et al., 2009). In passing it is interesting to note an observation made by Charles Darwin (Darwin, 1965/1872, p. 344), "Everyone feels blame more acutely than praise".

The greater valuation of loss over gain is understandable in evolutionary terms: the consequences of not being loss averse would be much worse than those of being gain prone. This principle of *loss aversion* converts within approach-avoidance theories to greater negative reinforcer sensitivity relative to positive reinforcer – *provided* the words 'negative reinforcer' are taken only in the sense of the valence of an objective event, independent of whether the contingency with which it is presented produces approach or avoidance.

We return later in this article to the methodological value of economic paradigms for the testing of approach-avoidance theories; but, for our current purposes, the key point to be derived from behavioural economics is that, on average, humans have distinct *valuation* sensitivities and losses are valued more than gains when holding approach/avoidance constant (Tom et al., 2007; Tversky and Kahneman, 1991). Given the within-subject use of money to demonstrate these effects, they cannot be due to differences in level of motivation, or of subtle differences of the kind that could give different results with supposedly 'matched' food and shock as reinforcers.

2.4 Motivation-Output – attraction and repulsion, not reward and punishment

Conceptually distinct from valuation is motivation-output; motivation is an inferred state and is indicated by actual behavioural output. If gain and loss sensitivity can differ, with approach or avoidance held constant, it follows that attractor and repulsor sensitivity can differ with gain or loss held constant (for evidence see Section 2.6). That is, attractor sensitivity operates with both gain and omission of loss, while repulsor sensitivity operates with both loss and omission of gain. The full implications of this distinction are not known for personality traits, and have not previously been considered – whether they are less important than implied here must await future empirical research. However, as the research evidence presented below indicates, there are grounds for assuming that they are sufficiently different to be taken seriously in future personality research.

From this we argue that the critical personality factors previously postulated by approach-avoidance theories should be termed 'attractor sensitivity' (approach) and 'repulsor sensitivity' (avoidance). On this view, response output systems control

approach and avoidance behaviours but these are distinct from stimulus input systems that process valuation. We have already argued that the economic concept of loss aversion demonstrates systematic differences between gain and loss valuation systems. However, approach results from both gain and loss omission, while avoidance results from both loss and gain omission. This requires that the simple *valuation* of gain or loss (independent of their contingencies) must operate orthogonally to attraction and repulsion. This does not imply that they must be orthogonal in terms of personality description and causation, but it points to this possibility.

This orthogonality means that the commonly used terms 'reward sensitivity' and 'punishment sensitivity' are ambiguous because they conflate valuation and action. This conflation is seen in the names of some commonly used personality measures, for example, the *Sensitivity to Punishment and Sensitivity to Reward Questionnaire* (SPSRQ, Torrubia et al., 2001). Some personality questionnaires focus on specific behaviours as opposed to reward and punishment per se, for example, the Carver and White (1994) *BIS/BAS Scales*, while others focuses more on evaluation in the context of reward and punishment: *General Reward and Punishment Expectancy Scales* (GRAPES, Ball and Zuckerman, 1990). In the context of the distinction we wish to emphasize, these scales are not interchangeable and have different construct validities, especially in terms of their internal factor structures and relations with broad measures of personality (For a review of this literature see, Torrubia et al., 2008). Reflecting this confusion, experimental results are often not consistent with the supposed relationship between the trait and state reward/punishment measures (e.g., Leue and Beauducel, 2008; Matthews and Gilliland, 1999; Pickering et al., 1997). Much debate in this literature has centred on which set of scales 'is best'. However, according to the position advanced here, these discrepancies may result from differences in the nature of these questionnaires - which

do not make a distinction between valuation and motivation aspects of approach-avoidance behaviour and may each combine these in different ways.

Thus, as we have seen, the previous analysis of 'reward' and 'punishment' sensitivities has focussed almost exclusively on discrete presentations of reinforcers to the exclusion of the value of their omission and so has confounded perception/valuation with motivation/action. However, omission of a positive reinforc*er* is negative reinforc*ement* (Amsel, 1992; Gray, 1987). It is, then, an open question as to how far psychometric measures of 'reward/punishment' traits actually relate to trait variation in valuation of explicit events and how far to trait variation in the strength of the action tendency that they produce. As discussed above, the literature suggests that such differences may be important. As we will see in section 2.6, these can be separated experimentally but they have generally not been so separated in the past.

2.5 Combining perception/valuation and motivation/action

In the two previous sections, we have discussed gain and loss as orthogonal to approach and avoidance, with the link between valuation and action being provided by contingency. This combination of valuation processes, contingencies and motivation processes, and the resultant potential sources of personality sensitivity, is shown diagrammatically in Figure 2. Actual stimuli (or their memory in the case of omission contingencies) are first evaluated in terms of their current exchange rate/drive level; this will then activate a central positive and/or negative valuation (Figure 1) that will determine, among other things, preference between two similar-valence alternatives. In our original discussion of loss aversion, the exchange-rate-based step that initiates valuation would have delivered equal $GAIN and $LOSS values. That is, amount interacts with exchange rate to generate the neural code for 'External value' in Figure 2.

This source-specific value is thus converted to a general positive or negative valuation, which provides the basis for choice between qualitatively different alternatives. Both of these general valuations has its own trait sensitivity. "Loss aversion", here, is the result of a greater population average (trait) negative valuation sensitivity *relative* to positive valuation sensitivity with the exchange rate of external money being, necessarily, the same in both cases. Conversely, the anhedonia that is a feature of some mental disorders (Treadway and Zald, 2011) would be seen as a reduction in trait positive valuation sensitivity.

----- Figure 2 about here -----

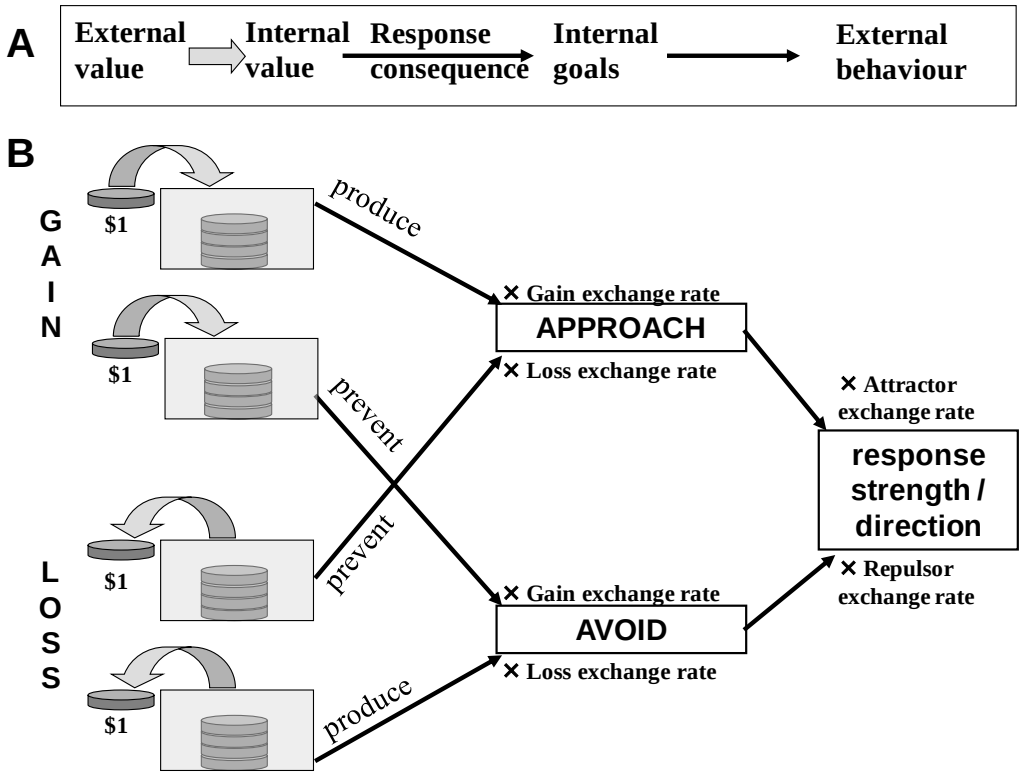

Figure 2. The combination of valuation and operant factors that determines response strength and direction. Items with a specific external value ($1) that can be gained or lost are represented by a particular internal amount that will depend on the exchange rate (or the level of "hunger") for

the item (see Figure 1). In this example all inputs are $1 and so exchange rate is ignored. The internal value that drives decisions and the intensity of action also depends on whether the item is gained or lost. Economic analysis has shown that the same external value generally has a greater effect if it is a loss (×Loss exchange rate) than if it is a gain (×Gain exchange rate). The effect of this internal valuation on behaviour then depends on the consequences of responding. Gain production and loss prevention activate approach; loss production and gain prevention activate avoidance. Concurrent APPROACH and AVOID tendencies are then integrated to determine the direction and strength of responding. A fixed internal value of approach and avoidance will have different effects on response strength (×Attractor exchange rate; × Repulsor exchange rate) that depends both on factors of reinforcement sensitivity and on the distance from the goal that will be achieved by responding. (Approach and avoidance have different goal gradients, see Section 2.8.)

Once the stimulus (or the memory of the stimulus in the case of omission contingencies) has been evaluated the direction of action is determined by the expected contingency of the action. An increase in positive value, or reduction in (or omission of expected) negative value (e.g. a reduction in chronic pain), leads to approach; and an increase in negative value, or reduction in (or omission of expected) positive value, leads to avoidance. The translation of this final, multiply adjusted, value into the strength of action is dependent on distinct approach and avoidance sensitivities.

2.6 Evidence for motivation as distinct from valuation in humans

The theoretical analysis we have just provided requires linking to experiment, both to provide evidence that a two stage theory is required and to provide a means of objectively assessing the postulated sensitivities in which personality theorists are interested. To assess attractor and repulsor sensitivity as general personality factors, the first requirement is to provide measures that, at the state level, eliminate the individual

exchange rates of specific motivational stimuli as well as the more general exchange rates that give rise to loss aversion.

Specific exchange rates result from variations in value due to, for example, variation in the current deprivational state of the animal for any particular class of appetitive or aversive item. As noted earlier, while these effects can seem difficult to match in the rodent paradigms that provide the neural foundations of approach-avoidance theories, it can be done simply in the more modern environment of behavioural economics. Money allows the easy manipulation not only of presentation and omission of the same motivationally significant item (such as $1) in opposition to each other but also the manipulation of whether the presented event is positive (gain) or negative (loss from an existing store). On this basis, the strength of reaction to these equated exchange-rate-constant values may be determined.

The more general exchange rates affecting attractors and repulsors result from the difference in valuation of gain and loss in absolute terms. To assess approach and avoidance while controlling for this valuation bias, Hall et al. (2011) administered a two phase task. In the first phase, humans started with $0 and could move a mouse to a target to gain money but with a 50:50 risk of losing money. Non-responding resulted in no gain or loss. The gain and loss values were fixed and known for blocks of trials during which response speed was measured and where all possible combinations of gain and loss were tested across blocks. The second phase was basically the same except that the same participants started with a set number of dollars and then each click prevented the loss of money but with a 50:50 risk of preventing the gain of money. (Note that from a ‘rational economic’ perspective there is no difference between this second phase and the first. A response that produces a gain of $1 and one that prevents the loss of $1 have identical consequences on take-home amount and so should be rationally valued equally.) Figure

3A shows the variation in speed with changes in attractor and repulsor value averaged across these two conditions (i.e. averaging across obtaining a dollar gain and preventing a dollar loss – eliminating, statistically the effects of loss aversion, see below). The procedure and measures are fundamentally similar to those used to test rats in runways.

----- Figure 3 here -----

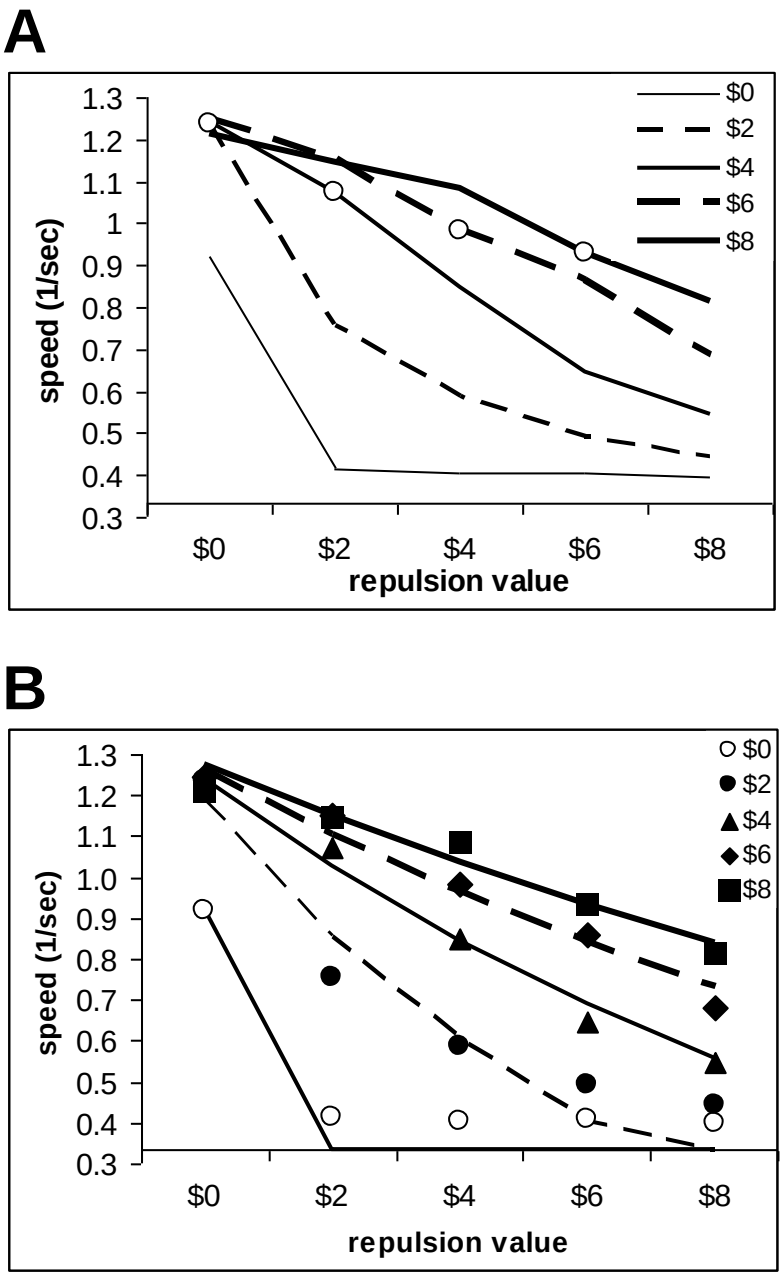

Figure 3. A. Observed speeds resulting from the combination of a specific dollar value (gain averaged with omission of loss – separate values plotted as separate curves) for the production of a response with a specific dollar value (loss averaged with the omission of gain – repulsor value, X axis) for the inhibition of the response. The probability for gain or loss on any particular trial was equal. Open circles indicate the point on each curve at which the net value averages to $1. B. The same data represented as point values with the curves resulting from the optimised fitted functions based on previous animal behaviour analysis (see text).

Important points to note about these results are that: (1) a net zero dollar value for the making of a response does not result in zero response (i.e. minimal speed); there appears to be an intrinsic value to responding even for no dollars and this outweighs the small intrinsic response cost that must also exist; (2) a net $1 difference between attractor and repulsor values produces speeds that depend on their absolute values rather than having a fixed "$1" effect; and (3) the interaction of attractor and repulsor values produces a curvilinear relationship (see especially the curve for a net $2 attractor value).

The observed curves are what would be expected from previous behavioural analysis of the variation in the speed of pigeon responding with variation in value (De Villiers, 1977; Killeen, 1994). The following function was derived from this literature and fitted to the data (Figure 3B).

$$\text{speed} = k \frac{A(a) - R(r)}{A(a) + R(r)} \quad \ldots\ldots\ldots\ldots \quad \text{Equation 1}$$

Where

$a = a_{intrinsic} + a_{extrinsic}$ = total attractor value for measured behaviour

$r = r_{intrinsic} + r_{extrinsic}$ = total repulsor value for measured behaviour

k = nominal maximum speed

A = Attractor sensitivity

R = Repulsor sensitivity

Extrinsic values are those imposed by the experimenter and are known. The two intrinsic values ($a_{intrinsic}$, $r_{intrinsic}$) and the parameters k, A and R must be calculated by a least squares fit to the data. This fit accounted for 98% of the variance in the averaged data and individual participant fits accounted for 89% on average of the individual data (range 58-96%). The goodness of these fits shows good generalisation of the previous animal analysis to this human task.

The critical result was that attractor sensitivity was always greater than repulsor (A:R ranging from 1.7 to 5.4 times the strength), thus the attractor and repulsor systems clearly have different sensitivities. It should also be noted that the variation in the ratio from 1.7 to 5.4 also suggests that there is variation in individual sensitivities. But, in typical animal experiments, approach and avoidance gradients are different (see section 2.8) and so the relative strength of the observed effects of an attractor and repulsor will change depending on distance from a goal – with a cross-over at intermediate distances. In this case, it could be that the ultimate goal (receipt of money at the end of the experiment) is relatively distant. We should also note that attraction being greater than repulsion would not be consistent with repulsion being directly linked to loss (given demonstrations of loss aversion) but, of course, any effect of loss *per se* was averaged out before fitting the curves (see below for the assessment of loss:gain independent of attraction:repulsion).

Perhaps the most important point for practical purposes is that, at least with these measures, attractor variance can only be assessed in the presence of some repulsor variance (or speed immediately goes to asymptote). Although the results are evidence for the requirement for a two-stage theory, and for their being differences in attractor and

repulsor sensitivity, they do not allow direct measurement of the individual sensitivities. We consider the implications of this for personality assessment later in this article.

2.7 Evidence for valuation as distinct from motivation in humans

While gain/loss sensitivity differences are well established in the economic literature, if we are to argue for a two-stage model it is important to demonstrate this within the same paradigm as we have demonstrated attractor/repulsor differences. Hall et al (2011) also assessed loss aversion, independent of attractor/repulsor differences, by calculating the *difference* between the use of gain and omission of loss to promote responding, and the *difference* between the use of omission of gain and presentation of loss to inhibit responding (i.e. the differences between the pairs of curves that were averaged to produce Figure 3). This difference (Figure 4) does, indeed, demonstrate loss aversion. That is, variations in speed are more extreme when omission of loss is used to generate responding than presentation of gain.

It should be noted here that our problem in resolving absolute attractor and repulsor sensitivities recurs with loss aversion. That is, loss aversion is measured relative to gain and is, in that sense, simply a ratio.

----- Figure 4 here -----

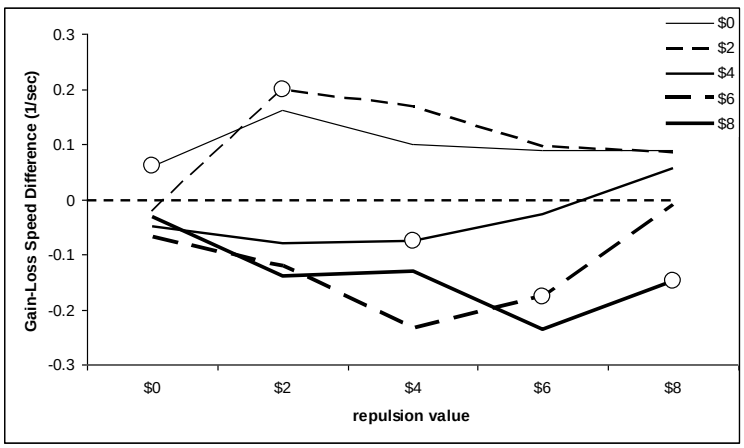

Figure 4. Observed differences in speed between gain and loss manipulations for the combination of a specific attraction value for the response with a particular repulsion value. Attraction dollar value is gain averaged with omission of loss (separate values plotted as separate curves). Repulsion dollar value is loss averaged with the omission of gain and is plotted on the X axis. The probability of gain or loss on any particular trial was equal. Open circles indicate the point on each curve at which the attraction and repulsion values are equal.

2.8 Attraction versus repulsion: subtraction, gradients and direction

We have so far discussed factors for which insufficiently tight definitions, or conflation of distinct meanings, will have previously led to difficulties in the testing of approach-avoidance theories. In this section, we deal with a number of other known parametric issues that will also be important for the generation and testing of quantitative predictions. That is, to apply even a simple binary approach/avoidance model to specific experiments, one needs to take into account some details of how attraction and repulsion operate and interact. A wealth of animal experiments have shown that attractors and repulsors of a particular value do not produce fixed behaviour nor, when they are co-activated, do they produce independent, or linear, effects on approach and avoidance tendencies (Mackintosh, 1974; Millenson and Leslie, 1979).

The most obvious interaction of attraction and repulsion is a subtractive effect on choice behaviour (see, Boureau and Dayan, 2010). If, when making a choice based on attraction, one of the alternatives is also moderately repulsive, the average rat or human will decrease their tendency to make that choice. Likewise, addition of attraction will reduce the tendency to make a choice that avoids repulsion.

What is less obvious is that co-activation of attraction and repulsion produces an additive effect on arousal. That is, while the probability of making a choice may be decreased when the opposing motivation is added, the intensity with which the

behaviour is emitted can be increased. The interaction of this increased arousal with decreased response probability can give rise to a variety of non-linear changes in observed behaviour of which behavioural contrast and peak shift have been subjected to the most detailed analysis in terms of the implied interactions between the approach and avoidance systems (Gray and Smith, 1969). One experimental consequence is that the addition of a mild repulsor can, paradoxically, invigorate ongoing attractor-controlled approach behaviour.

Early experiments analysing the speed of running in a runway also showed that the motivational value of attractors and repulsors varied with distance from a goal box – and that this goal gradient is steeper for repulsors than attractors (Gray, 1987; Miller, 1959). However, few experiments have been designed in such a way as to allow extraction of the relevant gradients or estimation of the subject's position along them. While the analysis of gradients in the animal literature has focussed on spatial distance from a goal, such gradients also clearly operate in relation to time in, for example, the acceleration of responding during the interval of a fixed interval schedule (Zeiler, 1977) and, within the human literature, in delay discounting (Kurth-Nelson and Redish, 2010).

The combination of the subtractive interaction between approach and avoidance tendencies, and the difference in their goal gradients, explains the typical behaviour when faced with the combination of an attractor and a repulsor as the outcome of action. The typical rat in a runway (or human in more complex, including social, situations) will initially approach the location at which both positive and negative consequences are available. At this initial, long distance, approach tendencies are stronger because their gradient is shallower. The closer to the location, the slower will be approach (since the strength of the avoidance tendency is increasing faster and is subtracting from approach);

until, if the repulsion is sufficiently strong, approach will cease before the location is reached and will be replaced by dithering and displacement activities, such as grooming.

2.9 Evidence for specific goal conflict processing in humans

We need to add goal conflict to the simple approach tendencies, simple avoidance tendencies and the symmetrical tendency of moderate levels of one to subtract from the other that we have considered so far. When there is significant approach-avoidance goal conflict, i.e. when approach and avoidance tendencies are both strong[1] and in relatively balanced opposition, then a third system is activated (Gray, 1977). This third system creates a need for a pure measure of its activation that is not contaminated, as all motor behaviour must be in a conflict situation, by interactions between pure approach and pure avoidance.

There is evidence that a specific component of EEG theta rhythm can provide a biomarker for goal conflict processing. According to Gray and McNaughton (2000), theta rhythm is important for the processing of conflict by the hippocampus and for its interaction with goal processing areas (including prefrontal cortex). It follows that, when the hippocampus is processing conflict, theta encoded output will pass from it to areas, such a prefrontal cortex, and so enhance any ongoing theta rhythms in those areas – and possibly entrain them as well (Young and McNaughton, 2009).

The occurrence of theta in the EEG has already been linked to anxious rumination (Andersen et al., 2009) and to goal conflict (Moore et al., 2006), and is known to be the EEG frequency band that best separates psychometrically defined low and high BIS individuals (Moore et al., 2012). However, cortical theta clearly represents

[1] When reinforcement values are small (failing to produce significant emotional involvement in responding) then goal conflict, as indexed by hippocampal lesions, is not engaged Okaichi and Okaichi (1994). This explains the lack of clear conflict effects in the Hall et al (2011) experiment.

a variety of processes and frontal midline theta, in particular, has the opposite pharmacology and personality correlates from those that would be expected of hippocampally-related theta (Mitchell et al., 2008). It is important when attempting to demonstrate conflict-related theta, then, to rule out effects of arousal, working memory, etc., as well as any effects of motivational stimuli linked simply to attraction and repulsion.

There is, however, a straightforward method of detecting activity in the conflict system. The requirement is for three experimental conditions: one generally eliciting approach; one eliciting avoidance; and one intermediate. Except for changes in value, all other aspects of the task should be the same across the conditions. The prediction is that, relative to both the net gain and the net loss conditions, the intermediate condition will result in increased theta power in areas such as the prefrontal cortex that are engaged in the control of the relevant behaviour.

Experiments of this type have already been conducted. Neo (2008) recruited human participants from a student job search pool; they volunteered for casual labour in exchange for cash amounts close to the minimum wage and so were likely to be motivated by money. EEG was taken while they performed a simple choice task in three different conditions. In all conditions there was a 50:50 probability of gaining 10c for pressing a left key and there were no monetary consequence for pressing the right key. Across the three conditions the other 50:50 alternatives were: losses of 0c (net average value +5c); 10c (net average value 0); and 20c (net average value -5c). The level of gain was, thus, constant across the conditions; and the level of loss (and the tendency to avoid the left key) increased steadily across conditions. However, the level of conflict in terms of gain-loss balance was greatest in the intermediate condition.

The results of Neo (2008) show that conflict-specific theta (that is greater theta power in the net 0c condition compared to the average of the +5c and -5c conditions) was observed most at right frontal and left posterior sites for both 4-5 Hz and 6-7 Hz theta. Given the results of Hall et al. (2011), it is important to note that all three conditions were equally ambiguous, all had equal risk (in terms of the probabilities of the outcomes) and loss was greatest with -5c net. The observed effect appears genuinely specific to conflict and cannot be attributed to any of the three classical neuroeconomic forms of aversion. The largest conflict effect was observed over the left temporal lobe (T3, 4-5Hz) and, across participants, the size of this effect was positively correlated with avoidance of left clicks (i.e. greater T3 theta predicted increased conflict-specific aversion. These results show the feasibility of using theta as a measure of state conflict processing, uncontaminated by activity in the approach and avoidance systems.

3 AN UPDATED STATE REINFOR*CER* THEORY

3.1 Reinforcers as a basis for a state theory of approach and avoidance

In the previous section, we discussed issues that we believe must be taken into account by any approach-avoidance theory of human personality. These issues are driven largely by experimental data and do not, in and of themselves, entail any particular theoretical integration. But to proceed to a specific theory of traits, it is important to have a coherent theory of the states for which trait factors provide consistent biases. Critically, trait factors express their effects on behaviour through state systems and their effects can only be properly explained in the context of an explicit state theory.

The state theory described in this section is based on that developed by Gray (1975, 1982) and recently modified and extended by us, and includes a quite detailed neurology that will not be discussed here (Gray and McNaughton 2000; McNaughton and Corr 2004, 2008b). The operation of the basic systems previously described is, here, further extended to an explicit two-stage model. We add valuation as a distinct input stage linked to a distinct goal-processing output stage and, in addition to the behavioural evidence we have discussed, there is neural evidence for distinct valuation and motivation mechanisms (e.g., Monosov and Hikosaka, 2012).

Despite the apparent complexity of its internal constructs and stages, the theory retains an unchanging, behaviourist, bedrock. Each aspect of each stage can be tested for its effects on behaviour or directly recorded neural activity (see Section 2) and for its pharmacological sensitivities. This explicit linking to behaviour and the nervous system is important because it provides an unambiguous, non-linguistic, set of measures to which all approach-avoidance theories can be linked and tested on an equal basis. That said, our insistence on behaviour or neuronal activity as evidence for inferred processes goes hand in hand with the view that even the lowly rat is driven by cognitions (e.g. goals, see below). Indeed, although superficially a paradox, it can be argued that behaviour analysis is the optimal means of assessing changes in the cognitive structures (Dickinson, 1980; McNaughton and Corr, 2008a) that underlie approach-avoidance behaviours

The output stage of the model retains the classic approach-avoidance assumption of two fundamental classes of *discrete, concrete,* motivating situation: positive and negative. These are *innate* attractors and repulsors (see Section 2), but it remains the case that making a stimulus (especially a secondary reinforcing stimulus) contingent on a response baseline and then seeing whether the baseline behaviour increases in frequency,

decreases, or does not change, is the simplest way of classifying that stimulus. To reflect this change in emphasis, we propose that the modified form of RST that we outline below be renamed *Reinforcer* (rather than *Reinforcement*) Sensitivity Theory. This captures the fact that both innate and acquired motivational stimuli (and the omission of expected stimuli) can be categorised into two fundamental classes based on whether they *are* positive or negative reinforcers *if* they are used in a conditioning paradigm.

The input stage of the model borrows wholesale from the work of behavioural economists on valuation. It proposes separate valuation systems for negative and for positive reinforcers – with a consistent trait difference between these generating loss aversion. Importantly, it also proposes that valuation is a potential confounding factor in the assessment of attraction and repulsion. The proposed input stage does not yet explicitly include risk aversion and ambiguity aversion. For the moment, we have treated these as simple sources of negative stimulus input, differing from others in the same way as do shock and cold. But they may require more complex treatment.

The core of the theory, then, is a combination of distinct gain and loss valuation input systems with approach and avoidance motivation output systems; with valuation and motivation orthogonalised by presentation/omission contingencies. We believe the existing data require such a structure of any theory of approach and avoidance.

3.2 FFFS and BIS: defensive direction

The key feature of our state RST, in comparison with more basic approach/avoidance theories, is that it postulates two (Gray, 1967, 1977; Miller, 1959) quite distinct avoidance systems – one for simple active avoidance and one for approach-avoidance conflict (passive avoidance). It identifies active avoidance with fear and a

Fight, Flight, Freeze System (FFFS), and it identifies approach-avoidance conflict with anxiety and a Behavioural Inhibition System (BIS), in common with the earlier versions of the state theory on which all versions of RST are based (Gray and McNaughton, 2000). By analogy with the concept of 'defensive distance' (which accounts for detailed variation in the nature of defensive responses with variation in the perceived level of threat and so the neural level of processing), the distinction between the FFFS and BIS can be seen as one of 'defensive direction' That is, the FFFS controls behaviours that have evolved to remove the animal from danger, while the BIS controls behaviours that have evolved to allow the animal to (cautiously) approach danger (Gray and McNaughton 2000; McNaughton and Corr 2004, 2008b).

The concept of defensive direction provides a single organising principle to define inputs to the BIS, whereas in 1982 Gray provided an *ad hoc* list. Importantly, it treats innate and acquired reactions equally (Blanchard et al., 2011) and the BIS, so defined, is generally sensitive to the anxiolytic drugs that provide the gold standard for the theory (Gray and McNaughton 2000, Appendix 1). The FFFS, by contrast, is *relatively* insensitive to anxiolytic drugs (or doses) but is sensitive to panicolytic ones (Blanchard and Blanchard, 1990a; Blanchard and Blanchard, 1990b; Blanchard et al., 1997). While predominantly studied in experiments with innate fear, the FFFS must be taken to control avoidance of all aversive stimuli, including learned ones.

The pharmacological distinction between the FFFS and BIS is particularly important when we wish to link this fundamentally rodent-derived theory with past and present work on human disorders and the personality types that are risk factors for them (Andrews et al., 1990; Duggan et al., 1995; Rovner and Casten, 2001; Roy, 1999). What are generally referred to clinically as 'anxiety disorders' (American Psychiatric Association, 2000) include what are, in ethological terms, both disorders of fear (e.g.

panic, simple phobia) and disorders of anxiety (e.g. agoraphobia, social anxiety) (Sylvers et al., 2011). It is important support for the theory that the specific anxiolytic drugs that define the BIS are effective in (ethologically defined) anxiety disorders but not fear disorders, whereas the panicolytic drugs that affect fear directly in rodent ethological tests are generally effective with fear disorders. A simple test of any paradigm intended for use in assessing BIS sensitivity, then, is whether it is similarly affected by not only classical (e.g. benzodiazepine) but also novel (e.g. buspirone) anxiolytics – since these two classes of drug do not affect fear, and share no side effects (McNaughton, 2002). Equally importantly, our capacity to define the BIS in terms of specific drug receptors argues for endogenous ligands (Carboni et al., 1996; Kapczinski et al., 1994; Montagna et al., 1995; Polc, 1995), variation in which can provide a substrate for the postulated variations in trait sensitivity (Abadie et al., 1999; Hode et al., 2000; Lehmann et al., 2002). Previous research has pointed to the need to differentiate fear and anxiety in personality questionnaires (McNaughton and Corr, 2004; Smillie et al., 2006b) and this call has been extended to clinical conditions (Bijttebier et al., 2009).

According to McNaughton and Corr (2004), the current value of defensive distance determines the key locus of control within the FFFS, but is not determined solely by the nominal value of perceived environmental threat. If there is concurrent conflict between goals (e.g. between approach to and avoidance of the same place) then the BIS is activated. An important feature of this activation is that, in addition to a tendency to inhibit both ongoing avoidance and ongoing approach and to replace these with risk assessment, the BIS increases negative cognitive bias; that is, it increases attention to negative stimuli and also amplifies the existing avoidance tendencies operating on the FFFS (and thence BIS), essentially decreasing the current defensive distance.

In the simplest case, BIS activation merely renders the animal risk averse – causing it to choose the less dangerous of two alternatives; to choose to leave the current situation if it can; or to remain in safety until the risk has diminished. However, it can also generate risk assessment or exploration (behavioural outputs of the BIS), or internal memory scanning. If these determine that threat is no longer present then activation of the BIS is terminated and behavioural control may revert back to BAS-mediated approach or, if threat is confirmed, FFFS-mediated avoidance behaviour.

As already noted above, it is important to remember that, in simple English, 'behavioural inhibition', if this means a reduction in behaviour, is not necessarily dependent on the BIS. This may seem paradoxical. When attraction and repulsion are not approximately equal in value, they subtract symmetrically (Gray and Smith, 1969). Repulsion can then reduce responding to attraction and this reduction is not sensitive to anxiolytic drugs (McNaughton and Gray, 1983). Thus, in our updated RST, the BIS is a system that not only amplifies attention and arousal (as previously postulated by Gray) but also *amplifies* the existing inhibition of behaviour. This amplification, in contrast to any background inhibition, only occurs under conditions of goal conflict. Behavioural inhibition pure and simple can also occur in the absence of the BIS when the level of conflict is low as a result of low levels of motivation (Okaichi and Okaichi, 1994) as noted, but only in passing, by Gray and McNaughton (2000).

3.3 The BAS, wanting, and liking

Our treatment of the BAS, as a system that processes attractors, also requires careful distinction between highly motivated behaviour and more simple action. Analysis of the systems (each represented at the level of the cortex, striatum, pallidum,

subthalamus, nigra, and thalamus) that control the production of motor behaviour delineates a set of hierarchical systems that control, in parallel, the selection of motor acts, actions and goals, respectively (Haber and Calzavara, 2009; Haegelen et al., 2009), with a key feature of the control of goals being that it is model-based (Boureau and Dayan, 2010). Both at the cognitive level, and in terms of the limbic structures involved, the BAS, as a global approach system, is best seen as operating with goals as opposed to acts or actions (for a matching model of parallel act, action and goal inhibition systems, see Neo et al., 2011).

An important point, here, that follows from the analysis we provided in Section 2.8 is that, while the initial stages of simple avoidance learning will involve avoidance of the negative goal of danger, when an avoidance response becomes well learned it will involve active approach to the positive goal of safety. So, while initially the FFFS will be involved, later control will shift to the BAS. Act and action generation, therefore, involves the BAS first identifying an attractor (defined by the combination of value and contingency) as a positive goal and then the cortico-striatal-nigral-thalamic system selecting actions that lead to the goal in part by concurrently inhibiting alternative actions via a mechanism that is independent of the inhibition of goals.

Our two-stage view of the BAS requires a distinct valuation stage as input to it. Consistent with this view, areas associated with goal processing, such as the orbitofrontal cortex and ventral striatum, have been reported to code stimulus value (Kang et al., 2011). Neurones in the amygdala do this too (Jenison et al., 2011) – linking the somewhat cold sounding concept of ‘value’ to emotional response.

We have also accepted the behavioural economic distinction between gain and loss that gives rise to loss aversion and there are data that suggest that the amygdala can control gain independently of loss (Weller et al., 2007). Conversely, amygdala

involvement, in contrast to the ventral striatum, appears to be lost when outcomes do not immediately follow responses (Tom et al., 2007). While the magnitude of the ventral striatal response differentiates between reward in the form of money and reward in the form of cognitive feedback, nonetheless, the same circuits appear to be involved in both cases (Daniel and Pollmann, 2010; Kang et al., 2011). All these data are consistent with there being two distinct systems that convert environmental inputs to internal common currencies of gain and loss, respectively.

A strong case has been made (Schultz, 2006, p. 87) that variations in the level of response of these neural systems can be linked to "basic theoretical terms of reward and uncertainty, such as contiguity, contingency, prediction error, magnitude, probability, expected value and variance". These terms are current in behavioural and economic theory. Many of these items relate to variation in the estimation of value or are examples of complex aversive reinforcers and have not yet been included in the theory presented here. For this reason, it was an important design feature of the experiment described in section 2.9 that it held ambiguity, uncertainty and risk constant while varying only value; and so eliminated their influences as potential confounds. The existing literature already contains a considerable quantity of information that will help us to determine suitable anchors for the separate aspects of gain valuation and approach (Kable and Glimcher, 2007; Padoa-Schioppa, 2011; Rushworth and Behrens, 2008) that we theorise underpin personality factors.

However, experiments of the type we have already discussed in section 2.6 and 2.9 that explicitly separate valuation from motivation, and which exclude changes in ambiguity, uncertainty and risk, appear to be largely lacking – but, when carried out, show that approach and withdrawal have effects independent of affective valence (Thibodeau, 2010). Conversely, activity in the amygdala has been linked, in particular,

to the framing effect that generates loss aversion (De Martino et al., 2006) producing equivalent effects whether the behavioural result is approach or avoidance.

In addition to distinguishing valuation and motivation phases that lead to action, we also need to note that "reward contains distinguishable psychological or functional components – 'liking' (pleasure/palatability) and 'wanting' (appetite/incentive motivation)" (Berridge, 1996). 'Liking' in this sense is probably distinguishable from valuation. "Hedonic '*liking'* by itself is simply a triggered affective state – there is no object of desire or incentive target, and no motivation for reward" (Berridge, 2004, p. 190). Liking, in this sense, has its own neural circuitry (Berridge, 1996, 2004). Wanting, by contrast, can be seen as containing all the action-generating components of 'reward' without accompanying sensory pleasure. This is particularly obvious in many cases of compulsive drug-taking by addicts (Berridge, 2004); and in this context it is worth noting that hedonic eating and drug taking involve similar neural circuits and response patterns (Kenny, 2011). However, particularly in relation to hedonic eating, 'wanting' and 'needing' are thought to be distinct (Finlayson et al., 2007) and so 'wanting' may not capture all of the action-generating aspects of what one normally terms 'reward'.

Wanting, if this analysis is correct, should contain both valuation and motivation components in the sense that we have been using these terms. Valuation here is not synonymous with pleasure (conscious or unconscious) but rather reflects the first stage of a two-stage process that, via contingency, results in action which may then end with pleasure. However, in the same way as it has been difficult to separate liking from wanting with the normal paradigms in which 'rewards' are delivered, it is difficult to separate valuation from motivation. Experimenters seldom compare the effects of matched positive and negative reinforcers in factorial combination with their

presentation and their removal (as opposed to omission). Separating out these different processes would seem a good target for future research.

4 A TRAIT REINFOR*C**ER* SENSITIVITY THEORY

In section 2, we discussed issues that, we contend, any approach-avoidance theory of personality should take into account. In this section, we apply these issues in the construction of a theory that incorporates positive evaluation, negative evaluation, attraction, repulsion and conflict. Although some degree of testing of such a personality theory can be carried out without a detailed state theory, rigorous quantitative testing depends on the precise way in which the various input-output relations of the different components of the systems interact. That is, it must depend on a detailed state theory that delineates the nature of the state interactions that can affect the output used to test for trait constants.

4.1 From state theory to trait theory

We endorse the view that such state theories should be neurally anchored. But, beyond this trite statement, what are the implications for personality theory of, for example, the detailed neural architecture proposed (McNaughton and Corr, 2004) for state control systems? Well, this neurology is intended to account for complex behaviour, the details of psychiatric disorder, and variation in the specific detailed effects of different drugs. But this detail focuses on moment-to-moment behaviour, and, in turn, these short-term variations are controlled by changes in more global external factors, such as the level of threat in the current environment. The major question is:

How should this state representation translate to the trait level? Below, we offer an answer to this question, but it is only one of a number of possible answers. Putting aside the specific correctness of the details of our proposed theory, we believe that the general form it takes has important implications for any such attempt to translate state systems to trait processes and measurements.

To illustrate one specific example, in the case of the FFFS, with escape and active avoidance, the concept of 'defensive distance' translates into the level of fear experienced: the closer the aversive stimulus, the greater the state of fear (e.g. the fear of death is not so terrifying when it is at some unknown time in the future; if it were tomorrow then fear would be greater). But, in addition to actual distance, and importantly in the case of clinical illness, level of fear is determined by *sensitivity* to the stimuli involved, that is, to fearful*ness*: some people show high levels of fear at a distance which for most people evoke, if at all, mild fearfulness (e.g. with air travel). Crucial here is the fact that level of fear experienced reflects the particular defensive behaviour shown. For example, each rat shows specific behaviours at its own particular actual distances from a particular predator – but with the different defensive behaviours appearing in the same sequence for all rats. It is the stability of this *sequence* that provided the original justification for the Blanchard's original concept of 'defensive distance'. So, too, in humans, a similar sequence is observed: mild threat elicits avoidance and flight, higher threat freezing, and intense and immediate threat panic and fight.

But these individual differences in fearfulness cannot be specific to a particular neural level of the FFFS: it is fundamental to the concept of defensive distance that fearfulness is a multiplier that controls *which* level of the system is selected by any particular standardised external intensity of threat. A highly fearful person, or rat, will

perceive a greater threat and thus activate a different module in the defence hierarchy (so they might panic when a less fearful person would actively avoid).

Likewise, anxiolytic drugs have been shown to affect defensive distance rather than to have a consistent effect on specific defensive behaviours. For example, in the rodent, rearing occurs at intermediate defensive distances. A reduction from a high level of threat to intermediate threat produces rearing, while further reduction in threat eliminates rearing and replaces it with normal daily behaviour. Administering an anxiolytic drug acts as if it is reducing threat rather than consistently increasing or consistently decreasing rearing behaviour as such. When rearing is low because of high threat, the drug increases rearing; but when rearing is high because of intermediate threat, the drug decreases rearing (Blanchard et al., 1991; McNaughton, 1985; McNaughton et al., 1984).

We have focussed here on the defense system because it has been analysed in detail, but Gray and McNaughton (2000) were quite explicit that positive goal-directed action was as hierarchically organised within the BAS as was negative within the defense system (see their Figure 9.4). However, for the same reasons as with our considerations of the defense system, we believe that the hierarchical levels of these systems can largely be ignored when considering trait approach sensitivity.

The factors that personality researchers should be most interested in, we believe, must act at least as generally as does an anxiolytic drug. They must normally control broad classes of behaviour – they should not be specific to behaviours within a class. However, this position leaves open the possibility that there may exist additional more specific personality factors related to the sensitivity/activity of more specific modules of the hierarchy (e.g. panic and obsession), and these sensitivities may be especially relevant to specific clinical conditions.

In principle, then, one might be concerned only with approach, avoidance and approach-avoidance conflict; but a major change of emphasis in our new reinforc*er* sensitivity theory adds to these gain and loss (i.e. general factors controlling the detection and valuation of environmental stimuli). Gain and loss (coupled with appropriate contingencies) serve as inputs to attraction and repulsion systems (the BAS and FFFS, respectively). Thus, there is a cascade of effects from gain/loss valuation, interacting with contingency, to attraction/repulsion. Within this framework, repulsion and attraction are probably at the heart of what were previously called 'punishment' and 'reward' sensitivities – but where valuation was not considered a separate stage. As discussed above, at present these processes seem conflated in the different psychometric measures of reward/approach and punishment/avoidance. We contend that this addition of loss and gain valuation (input) systems to BAS, FFFS and BIS (output) disentangles the previously problematic categories of 'reward' and 'punishment'. The cost of cutting this Gordian Knot is an increased level of theoretical complexity but, it is to be hoped, it will be repaid by increased conceptual clarity and experimental precision; and certainly each element can, as we have seen above, be tested independently of the others.

4.2 The trait theory

Our proposed trait theory derives directly from the state theory described above by ascribing specific long term, trait, sensitivities to the operation of specific links in the state systems. These links are indicated in Figure 5 by stippled shading.

As summarised in Figure 5, the BAS and FFFS are primary affective systems responsible for approach and avoidance, respectively. They will be activated in isolation when an attractor (e.g. PosR+, NegR-; i.e. explicit positive reinforcer presentation,

explicit negative reinforcer omission) needs only to be approached, and a repulsor (e.g. NegR+, PosR-,) needs only to be avoided, respectively. In addition to the variations in motivational level and other factors that affect the value of specific reinforcers, the model includes general PosR/NegR sensitivity differences to accommodate the loss aversion (i.e. NegR>PosR) demonstrated by behavioural economics (Tversky and Kahneman, 1991) – also specifically exemplified in section 2.7.

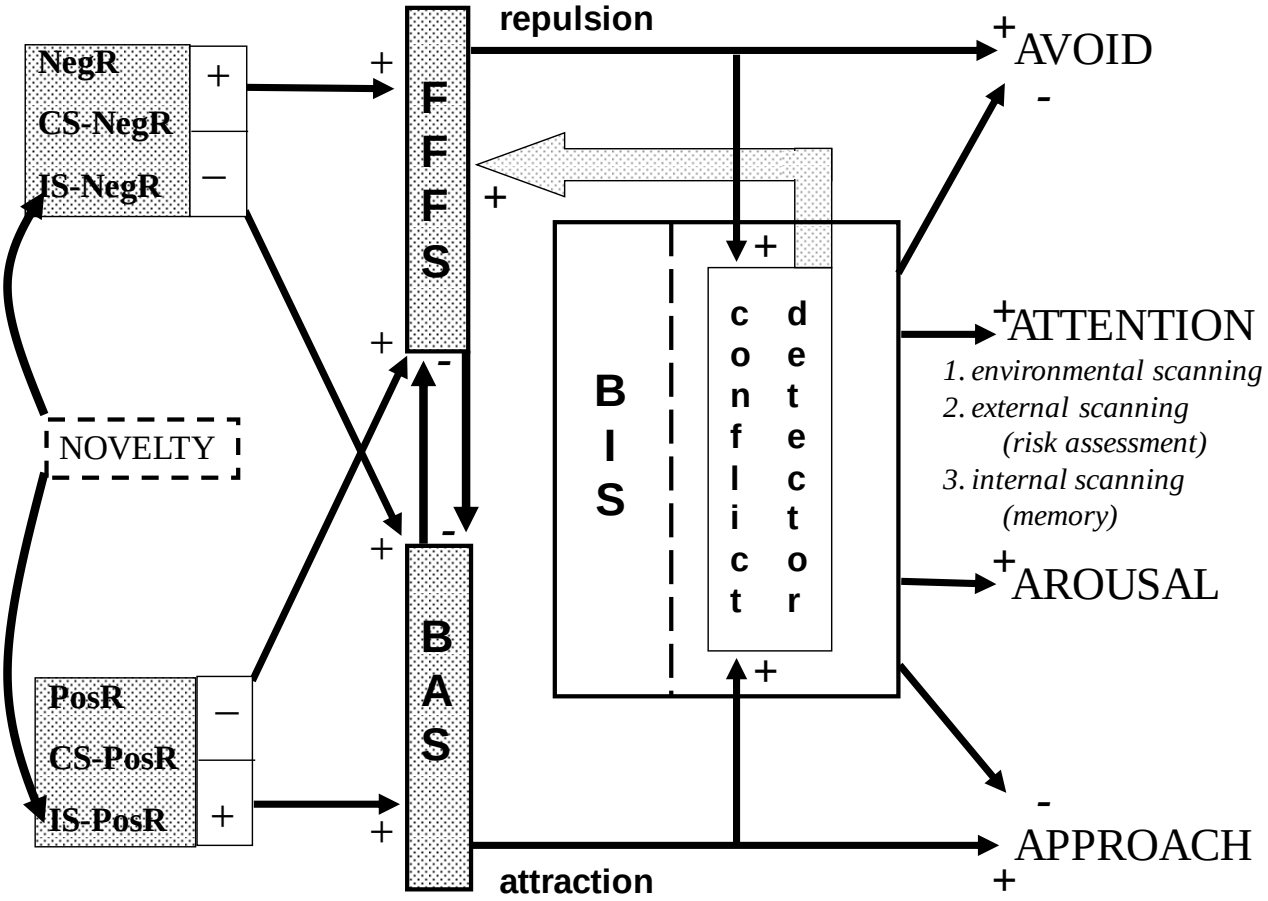

Figure 5. Overall relation of the BIS, FFFS and BAS – an updated model. To activate the BIS one must generate concurrent and approximately equal activation of the FFFS and the BAS, i.e. face the animal with an approach-avoidance conflict. Both simple approach and simple avoidance will then be inhibited and replaced with environmental scanning (in the form of altered attention), external scanning (risk assessment behaviour) and internal scanning of memory. Note that all of these scanning operations are aimed at detecting affectively negative information and involve a selective increase (stippled arrow) in the salience and value of aversive information. As a result, a secondary consequence of activation of the system is normally a shift of the balance between approach and avoidance tendencies in the direction of avoidance. However, when scanning determines that danger is absent the approach-avoidance conflict is resolved in favour of approach. The inputs to the system are classified in terms of the delivery (+) or omission (-) of

primary positive reinforcers (PosR) or primary negative reinforcers (NegR) or conditional stimuli (CS) or innate stimuli (IS) that predict such primary events. (Adapted from Gray and McNaughton, 2000.) As discussed earlier we see loss (i.e. removal of a positive reinforcer from an existing store) as a form of NegR, thus allowing for both Loss+ and Loss- (see section 2.6, 2.7). Specific cases of PosR and NegR will have their own individual exchange rates but, as discussed in the text, their effect will also be modulated by a general sensitivity factor that is different for the two classes of reinforcer. The stippled areas in the model are all points at which general personality factors could operate (see section 4).

----- Figure 5 about here -----

There are two features of this scheme that make testing via economic experiments particularly attractive. First, is that PosR+ and PosR- operate on the BAS and FFFS, respectively, so manipulation of just one specific reinforcer (PosR, in dollars, say) should be able to assess the relative sensitivity of the BAS and FFFS since the absolute value of PosR+ and PosR- can be made equal and, except for the change in direction, their exchange rates must be equal also. Second is that NegR can involve explicit loss from an existing store and so this expected loss can itself be experimentally omitted, generating NegR-. It follows that, if one compares addition of dollars to a store with removal of dollars from the store, one can compare PosR and NegR with the assurance that the specific exchange rate (including variation in drive) is matched.

It is possible, then, to assess the value of loss *relative to* gain and the value of an attractor (approach) *relative to* a repulsor (avoidance) using money. With a store of dollars, gain (addition of dollars), omission of expected gain, loss (removal of dollars), and omission of expected loss are all valued in the same currency. This allows the differences between PosR and NegR (whether presented or omitted) and of attraction and repulsion (averaged over gain and loss) can be assessed on an equal footing – as

demonstrated in section 2.6. It is possible to make similar arrangements with rats (e.g. with food being stolen from a store) but technically much more difficult to arrange clear parity between PosR and NegR.

In terms of general approach-avoidance behaviour, 5 separate basic sensitivities need to be considered: positive valuation and negative valuation (on the input side), and approach, avoidance and conflict (on the output side): 'motivation' is the inferred central state linking these two sets of processes. Assuming that these processes are independent at the trait level, our state tests must take into account (or counterbalance out) their likely interactions (e.g. where a variation in sensitivity to gain can result in increased approach, independent of approach sensitivity, *per se*).

To deal with this number of sensitivities, and the likely interactions between them, it is necessary to use a combination of neural measures with carefully selected, non-linguistic, behavioural paradigms. Both the measures and paradigms need to be driven by theory with strong *a priori* hypotheses; and this is why a specific trait theory is needed and specified below, as a starting point for experimental attack and refutation. In this context, neural measures have an advantage in that they can tap directly into each of the specific biological components that are the substrate of a theory with the minimum number of assumptions. Certainly, neural-behavioural relations uncovered to test one theory are immediately applicable to other theories. Given Smillie's (2008) extensive discussion of neuroscience paradigms, detailed discussion is not needed here.

In relation to research approaches in human participants, several key points are worthy of note: 1) neural imaging and EEG measures can selectively assess internal reactions specific to a reinforcer or class of reinforcers and, critically, can isolate stages of processing that are conflated in behavioural output (examples have already been given, above); 2) drugs can be used to target the key modulatory systems and so

challenge the involvement of particular neural or behavioural measures; and, especially when one wants to link states to traits (e.g., Perkins et al., 2009) – on this view, drugs may be mimicking endogenous compounds that supply the proximal basis for some traits, while genes can be viewed as a relatively fixed source of this intrinsic neurochemical variation; 3) molecular genetics can identify at least some aspects of some trait components of the state systems (e.g., Perkins et al., 2011).

Technology now allows for sophisticated designs that combine neural measures (e.g. fMRI and EEG), molecular genetics (e.g. candidate genes, or increasingly genome-wide scans) and psychometric measures of personality traits with carefully selected stimuli, of the type discussed above, to activate selective parts of the defensive and approach systems. A general theory, of the type presented here, would facilitate this research work.

4.3 Higher order factors: Neuroticism/worry

To identify specific trait sensitivities for each of a set of distinct neural approach-avoidance systems does not rule out the possibility that higher-order factors may mediate changes in more than one of the systems we have defined. In particular, our analysis calls for separate defensive trait sensitivities for FFFS-related fear/active avoidance (with avoidant personality disorder potentially representing an extreme) and BIS-related anxiety/passive avoidance/conflict (with generalised anxiety disorder representing an extreme). However, the two underlying neural systems are equally innervated by monoamine systems (serotonin and noradrenaline) and, if one sees drug treatment of psychiatric disorder as operating through something akin to a change in personality, one must note that the more general serotonergic drugs (i.e. drugs that are

not specific to 5HT1A receptors) treat fear and anxiety disorders equally well (not to mention also treating depression). In addition to specific active avoidance and conflict factors, then, there appears to be a higher order serotonergic/noradrenergic factor.

If we look among current scales for a possible example of such a superordinate factor, Eysenck's Neuroticism immediately obtrudes itself. Importantly, it appears to be an indiscriminate risk factor for the development of both fear and anxiety disorders (as well as depression) (Andrews et al., 1990). A risk factor for a set of things cannot be any one of those things itself and so a simple explanation of the findings is that neuroticism is a factor that, in the long term, or interacting with extreme events, can result in increases in trait anxiety and/or trait fear and/or trait depression (with extremes of each of these constituting disorder). Indeed, as we have argued in relation to the role of dopamine, the serotonergic factor may relate more to the modifiability of trait fear and trait anxiety than directly influencing their values. This suggestion is consistent with the apparent partial relationship between Eysenck's Neuroticism and Trait Anxiety (Spielberger et al., 1983), provided we presume that the latter is more closely measuring trait anxiety as we have defined it in our theory.

Following on from this view of neuroticism, it is worth taking a close look at one of its components: worry. The psychological state corresponding to what in the neuropsychological (but not economic) literature would be called risk assessment (including memory scanning) can perhaps be viewed by some as worry – if by this we mean simply the immediate perception of approach-avoidance conflict. However, if by 'worry' one means iterative rumination, this has not been studied in rats and was not explicitly dealt with even by the most recent detailed exposition of state RST (Gray and McNaughton, 2000). Critically, such ruminative worry appears to depend on a factor that is independent of anxiety in the most basic sense of the term (Meyer et al., 1990).

Further, worriers in the sense of those with a tendency to iterative rumination seem characterised by a general failure to control negative cognitive intrusions whether these relate to simple avoidance or to conflict (Borkovec et al., 1983). A tendency to worry is, therefore, a risk factor for anxiety disorders because such negative rumination can result in a failure to resolve the underlying conflict (e.g. by worry itself producing further conflict). This failure of ruminative control is also typical of disorders of simple avoidance, such as obsessive compulsive disorder, that are insensitive to the anxiolytic drugs that define the BIS but are sensitive to frontal cortical lesions (Powell, 1979). More work needs to be directed at clarifying the relationship between the FFFS/BIS and Neuroticism, as well as the role played by worry. Here, genetic approaches may be especially useful in delineating their structural properties, and neural and behavioural measures their process relationships.

4.4 Higher order factors: Extraversion

Previous work has superficially linked dopamine with ‘reward’ and extraversion. But, as we noted earlier, ‘reward’ in this literature conflates gain with approach and dopamine release is not reward-related as such. Extraversion is, therefore, likely to be a superfactor of the same type as neuroticism.

As well as distinguishing gain and approach we have also already made a distinction between ‘wanting’ and ‘liking’. Extraversion has two separable but correlated subfactors that emerge from factor analysis of many Extraversion facets (DeYoung et al., 2007). DeYoung (2010) has hypothesized that the two major subfactors within Extraversion may reflect the distinction between sensitivity of the BAS and sensitivity of a ‘pleasure system’ (PS). Likewise, the most popular supposed psychometric measure of

the BAS in personality studies, the Carver & White (1994) scale, has three separate sub-scales: *Drive*, *Reward Responsivity*, and *Fun Seeking*. Whereas Drive and Reward Responsivity both appear to characterize sensitivity to reward primarily, Fun Seeking appears to be equally related to impulsivity, and thus may not be as pure an indicator of BAS sensitivity (Smillie et al., 2006a; Wacker et al., 2012).

In sum, while variation in the dopamine system is clearly an important factor that must be taken into account by RST, the dopamine signal does not appear to equate with either gain or approach as such. If it is linked to extraversion, then, this may be a superordinate factor related to the modification of responding rather than representing either trait gain or trait approach. This would match the role we have attributed to the other monoamines and neuroticism in the previous section. This suggestion would be consistent with one of the several subfactors of extraversion or "BAS" scales being a specific measure of the BAS as defined here, while others could relate to gain or to pleasure.

4.5 Lower order factors

While we have argued, above, that personality theorists can deal with approach and avoidance systems with a 'lumping' strategy at the neural level, we need to add at least one caveat. There is reason to believe that there are at least two subordinate trait defense factors: obsessionality, likely linked to a subset of serotonergic receptor systems limited to areas like the cingulate cortex; and panic, likely linked to variations in the CCK system (Wang et al., 1998) and to local changes in the periaqueductal gray (Graeff, 1991). These are easily derived from the current state theory as trait variation within already-defined (McNaughton and Corr, 2004) specific modules of the avoidance system but should clearly be seen as additional to RST as it is conventionally framed.

4.6 Neuroscience Anchoring of Traits to State Systems

Having proposed the broad outlines of a trait theory, and after noting that much more experimental work is needed to test and develop this theory, the next question concerns the 'anchoring problem' that attends any attempt to relate trait models to state systems. This is an important matter for testing the proposals we have made above.

Factors recovered from questionnaires may reflect semantic rather than biological regularities and this is particularly the case if the items included in a questionnaire are chosen for their semantic content (e.g. asking people about presumed approach or avoidance behaviour) rather than for their links to presumed biological states (e.g. the sleep and weight items in the Beck Depression Inventory). Thus, causal theories of personality face a fundamental problem in identifying biological (and cognitive) systems that underlie personality factors. Most importantly, the use of factor analysis creates a problem in that it does not anchor the extracted factors within the multidimensional space that it derives (Block, 1995; Corr and McNaughton, 2008; Lykken, 1971). Therefore, factor analysis can provide only a preliminary guide to the biological processes underlying the most common trait variations in a population.

Given a) behavioural paradigms that can activate each of the differently valenced components of the different stages of processing, and b) concurrent EEG and/or fMRI measures, which allow stages of processing to be separated on the basis of both latency and neural location, we should be able to identify separate values for gain, loss, attraction and repulsion – and so anchor them (Gray et al., 2005; Reuter, 2008). fMRI is already being used to assess valuation (Trepel et al., 2005); and extreme cases of anhedonia may also be helpful here (for a review of anhedonia and depression, see Treadway and Zald, 2011). A good example of this form of approach is provided by Cunningham et al. (2010), who reported the association of fMRI-defined amygdala

reactions with two major facets of Neuroticism taken from the Five-Factor Model: Volatility and Withdrawal, which have previously been related to the FFFS and BIS, respectively (DeYoung et al., 2007). In the Cunningham et al. (2010) study, participants were presented with positive, negative, and neutral images and were required to approach (move perceptually closer to) or avoid (move perceptually farther away from) stimuli in different blocks of trials – this relates to the defensive direction hypothesis of RST (McNaughton and Corr, 2004). Results showed that higher scores on Volatility increased amygdala activation to negative stimuli (regardless of whether they were approached or avoided), while higher scores on Withdrawal increased amygdala activation to all approached stimuli (regardless of stimulus valence). A similar approach could be pursued to separate valuation from motivational outputs.

Once pure avoidance (FFFS-related) and pure approach (BAS-related) sensitivities have been estimated then their conflict may be measured by the EEG theta rhythm. The capacity to change theta rhythm is diagnostic of anxiolytic action, with at present no false positive and no false negatives (McNaughton et al., 2007). Neuropsychological theory (Gray and McNaughton, 2000) predicts that theta-encoded output from the hippocampus will invade other structures only when the hippocampus is producing functional output. The expected resultant behaviour-dependent phase-locking of theta rhythm between the hippocampus and prefrontal cortex has been demonstrated in rats (Young and McNaughton, 2009). Conflict-specific increases in frontal (or other cortical) theta power in the human EEG can, therefore, potentially, provide a pure measure of BIS activation (e.g. Andersen et al., 2009), and this rhythm has been shown to differentiate psychometrically-defined low and high BIS individuals (Moore et al, 2012).

However, some care is required here. Results such as those reported by Andersen et al. (2009), showing enhanced theta coherence during experimentally-induced anxious rumination, may provide support for the theory. But they do not meet the strict criterion of examining behaviourally-defined conflict in a situation where the influences of attraction and repulsion can be excluded. Previous research has not controlled for simple approach and avoidance before testing the effects of their conflict. This problem was addressed by Neo (2008, see section 2.12) who assessed theta-related conflict at different combined levels of attraction and repulsion and, critically, found no difference between the approach and avoidance conditions (in contrast to a difference between the average of these two and the intermediate conflict condition). In a similar way to the extraction of gain, loss, attraction and repulsion with EEG or fMRI, trait conflict could be extracted using theta as a state anchor for trait items in questionnaires or other measures.

This brief discussion highlights the necessity of theoretically-driven experimental designs to isolate specific components of valuation (input) and motivational (output) systems. To date, this has not been achieved, arguably because the theory on which such studies have been based has been insufficiently precise.

4.7 Anchoring trait measures – the value of drugs

In addition to the paradigmatic manipulations described for the generation of conflict-specific theta, we believe it will be important also to validate any state conflict measure such as theta with drugs before proceeding to use it as an anchor for questionnaire or other items already known to be linked to trait factors.

The key neurobiological aspect of the BIS theory is its derivation from the effects of anxiolytic drugs, which were only subsequently linked to hippocampal theta

rhythm and to conflict. Drugs are always messy instruments and, in relation to the BIS, the key requirement is to demonstrate common effects of both classical (e.g. benzodiazepine) and novel (serotonergic, e.g. buspirone or SSRIs) drugs. These distinct classes of anxiolytic have no common clinical side effects and not only affect clinical anxiety but also affect hippocampal theta rhythm (McNaughton et al., 2007). Critically, anxiolytic benzodiazepines and 5HT1A acting drugs, such as buspirone, affect the BIS but not FFFS. So, to assess BIS involvement in any neural or behavioural system we just need to show a common effect of the two classes of anxiolytic.

The key point is that a state measure of conflict, including theta, should only be accepted if it shows a *decrease* (ideally, for theta, in both frequency and amplitude) with *both* benzodiazepine and serotonergic anxiolytic drugs. This is in contrast to studies of frontal midline theta, which have shown *increases* with both classes of drug (for review, see section 6.3.1 in Mitchell et al., 2008). Indeed, conflict-related theta that appears at right frontal sites in the stop signal task (Neo et al., 2011) has been shown to be decreased by both classes of drug when other theta, in the same task, was increased (McNaughton et al, submitted).

There are a variety of reasons for seeing drugs as final crucial touchstones for tests of the state aspects of the theory and, potentially, genetics as the touchstone for traits. However, we emphasise behavioural and EEG (or imaging) methods above because, while drugs can validate a test, they operate on states and cannot assess the longer-term, personality-linked variation in the character that can then be assessed by the test they have validated. By contrast, imaging and EEG can provide amplitude values for a particular state response and these can be compared across people and so correlated with their scores on personality scales. We, thus, have to find tests (like those exemplified in section 2 above) with strong theoretical underpinnings to assess (in the

absence of drugs) personality variation (which, only once determined can we link to genetics).

In this discussion, we have focussed on conflict and non-panicolytic anxiolytic drugs. However, detailed neural theory also suggests that dopamine should be involved more in, and define sensitivity to, the BAS; serotonin (coupled with noradrenaline), both the FFFS and BIS. Likewise, paralleling the mutual opposition between the BAS and FFFS, it has been suggested there is mutual opposition between the dopamine and serotonin systems (Boureau and Dayan, 2010). However, we have already noted that the monoamines all have actions that span the systems we have delineated and so are unlikely to embody any one of the 5 specific RST sensitivities we have proposed.

5 CONCLUSION

5.1 Summary of main points

1. The concepts of 'reward sensitivity' and 'punishment sensitivity' at the core of a broad family of reinforcement-based personality theories need to be replaced. In the case of 'punishment', we have identified at least three neurally-distinct meanings of the term.
2. The most novel proposal of this work is that personality theory must take into account the *valuation* of positive (gain) and negative (loss) events and treat these as orthogonal to the *motivation* to approach or avoid. On this view 'loss aversion', as studied in economics, represents a systematic population difference between personality factors of gain and loss sensitivity.
3. Valuation must, then, be combined with a contingency of presentation or omission to generate an attractor or repulsor. Attractors and repulsors then

operate via distinct sensitivities (unrelated to the sensitivities of the two valuation systems) to activate the BAS and FFFS, respectively.

4. There are three (not two) fundamental systems controlling the output of motivated behaviour (whether innate or learned): attraction/approach (BAS), repulsion/avoidance (FFFS), and conflict resolution (BIS; this is responsible for inhibition of pre-potent behaviour and activation of threat assessment when there is a similar and concurrent activation of the BAS and FFFS that is producing emotional goal conflict).
5. We speculate that there may be five primary personality sensitivities related to reinforcers: positive evaluation, negative evaluation, attraction, repulsion and conflict. However, further empirical work is needed to determine if these separate processes are represented at the level of personality.
6. At present, there would appear to be general superordinate traits of 'neuroticism/emotionality' (linked to noradrenalin and serotonin) related to both the FFFS and BIS; and 'extraversion' (linked to dopamine) related to the BAS and some other aspects of positive affect. There also appear to exist smaller scale traits (e.g. obsessionality and panic proneness). We argue that a range of psychopathologies reflect the extreme ends of a normal distribution of one or more of these various (superordinate, RST, subordinate) factors.
7. We do not assume that the fundamental biological entities controlling these traits are necessarily completely independent. However, even if the fundamental trait variables are orthogonal, our analysis of the state systems indicates that there will likely be interaction between them both concurrently (as with the subtraction of approach tendencies from avoidance tendencies) and sequentially (as with the capacity for neuroticism to act as a risk factor for trait anxiety and trait fear).

Assessing systems that, *ex hypothesi*, will interact to generate any specific behaviour creates special problems for experimental design. Examples of some initial empirical demonstrations of system separation and specific measurement were provided.

8. These proposals require experimentation to test their validity; and we suggest that such testing would be best attempted via a combination of behavioural analysis, neuroscientific assays and both intra- and inter-individual personality study.

5.2 Final words

We have presented a review of issues that, we believe, are important for approach-avoidance theories in general and Reinforcement Sensitivity Theory (RST) in particular. We follow Smillie et al (2006b) in believing that revised state RST holds important implications for how the personality traits associated with its systems should be measured. Our clarification and elaboration of state RST highlights the problems that must be addressed in translating to a trait RST model, which we hope is a step towards resolving these problems and, ultimately, integrating the entire family of approach-avoidance personality theories. Specifically, we have called attention to three major issues: (1) the relevance of findings from behavioural economics, relating to gain evaluation and loss aversion; (2) the conflation of perception-valuation and motivation-action and thus confusion surrounding the use of the terms ‘reward’ and, even more so, ‘punishment’; and, (3) the common role played by innate and conditioned *reinforcers*, and so the inappropriate status of the term *reinforcement* in the family of personality theories that have approach and avoidance systems at their core.

This article concludes with the suggestion that in the form in which we have now cast it, the Reinforc*ement* Sensitivity Theory (RST) would be better termed the Reinforc*er* Sensitivity Theory of personality, representing all five systems we have identified, as contrasted with the revised three systems (i.e. FFFS, BIS, BAS; Gray and McNaughton, 2000) and the classic two systems (BIS/BAS; Gray, 1982) previous versions. For the sake of clarity, we suggest that in future writings these are respectively differentiated as: RST-5, RST-3, RST-2.

Acknowledgements

We each owe independent debts of gratitude to the late Professor Jeffrey Gray for stimulating the ideas that have come together in this article. He also read a much earlier draft of the manuscript of which this is now an extract and extension; and he provided critical insights into its form and content. Professor Gerald Matthews also provided valuable comment on the same earlier draft. We are most grateful to referees of previous versions of this paper. Their careful and perceptive criticisms and extensive suggestions for improvement helped us to refine the goals and content of the final article.

REFERENCES

Abadie, P., Boulenger, J.P., Benali, K., Barré, L., Zarifian, E., Baron, J.C., 1999. Relationships between trait and state anxiety and the central benzodiazepine receptor: a PET study. Eur. J. Neurosci. 11, 1470-1478.

American Psychiatric Association, 2000. Diagnostic and statistical manual of mental disorders: DSM-IV-TR. Amer Psychiatric Pub Inc.

Amsel, A., 1992. Frustration theory: an analysis of dispositional learning and memory. Cambridge University Press, Cambridge.

Andersen, S.B., Moore, R.A., Venables, L., Corr, P.J., 2009. Electrophysiological correlates of anxious rumination. Int. J. Psychophysiol. 71, 156-169.

Andrews, G., Stewart, G., Morris-Yates, A., Holt, P., Henderson, S., 1990. Evidence for a general neurotic syndrome. Br. J. Psychiatry 157, 6-12.

Ball, S., Zuckerman, M., 1990. Sensation seeking, Eysenck's personality dimensions and reinforcement sensitivity in concept formation. Personality and Individual Differences 11, 343–345.

Berridge, K.C., 1996. Food reward: brain substrates of wanting and liking. Neurosci. Biobehav. Rev. 20, 1-25.

Berridge, K.C., 2004. Motivation concepts in behavioral neuroscience. Physiol. Behav. 81, 179-209.

Bijttebier, P., Beck, I., Claes, L., Vandereycken, W., 2009. Gray's Reinforcement Sensitivity Theory as a framework for research on personality-psychopathology associations. Clin. Psychol. Rev. 29, 421-430.

Blanchard, D.C., Blanchard, R.J., 1990a. Effects of ethanol, benzodiazepines and serotonin compounds on ethopharmacological models of anxiety, in: McNaughton, N., Andrews, G. (Eds.), Anxiety. University of Otago Press, Dunedin, pp. 188-200.

Blanchard, D.C., Blanchard, R.J., Rodgers, R.J., 1991. Risk assessment and animal models of anxiety. Animal models in Psychopharmacology, 117-134.

Blanchard, D.C., Griebel, G., Pobbe, R., Blanchard, R.J., 2011. Risk assessment as an evolved threat detection and analysis process. Neurosci. Biobehav. Rev. 35, 991-998.

Blanchard, R.J., Blanchard, D.C., 1972. Effects of hippocampal lesions on the rat's reaction to a cat. J. Comp. Physiol. Psychol. 78, 77-82.

Blanchard, R.J., Blanchard, D.C., 1990b. An ethoexperimental analysis of defense, fear and anxiety, in: McNaughton, N., Andrews, G. (Eds.), Anxiety. Otago University Press, Dunedin, pp. 124-133.

Blanchard, R.J., Griebel, G., Henrie, J.A., Blanchard, D.C., 1997. Differentiation of anxiolytic and panicolytic drugs by effects on rat and mouse defense test batteries. Neurosci. Biobehav. Rev. 21, 783-789.

Block, J., 1995. A contrarian view of the five-factor approach to personality description. Psychol. Bull. 117, 187-215.

Borghans, L., Duckworth, A.L., Heckman, J.J., ter Weel, B., 2008. The economics and psychology of personality traits. . J. Hum. Resour. XLIII, 972-1059.

Borkovec, T.D., Robinson, E., Pruzinsky, T., DePree, J.A., 1983. Preliminary exploration of worry: some characteristics and processes. Behav. Res. Ther. 21, 9-16.

Boureau, Y.-L., Dayan, P., 2010. Opponency revisited: competition and cooperation between dopamine and serotonin. Neuropsychopharmacology Reviews 36, 74-97.

Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O., 2010. Dopamine in motivational control: rewarding, aversive, and alerting. Neuron 68, 815-834.
Carboni, E., Wieland, S., Lan, N.C., Gee, K.W., 1996. Anxiolytic properties of endogenously occurring pregnanediols in two rodent models of anxiety. Psychopharmacology (Berl). 126, 173-178.
Carver, C.S., 2004. Negative affects deriving from the behavioural approach system. Emotion 4, 3-22.
Carver, C.S., 2008. Two distinct bases of inhibition of behavior: viewing biological phenomena through the lens of psychological theory. European Journal of Personality 22, 388-390.
Carver, C.S., Harmon-Jones, E., 2009. Anger is an approach-related affect: evidence and implications. Psychol. Bull. 135, 183-204.
Carver, C.S., Johnson, S.L., Joormann, J., 2008. Serotonergic function, two-mode models of self-regulation, and vulnerability to depression: what depression has in common with impulsive aggression. Psychol. Bull. 134, 912-943.
Carver, C.S., White, T.L., 1994. Behavioral inhibition, behavioral activation, and the experience of affect: the BIS/BAS scales. J. Pers. Soc. Psychol. 67, 319-333.
Cloninger, C.R., 1986. A unified biosocial theory of personality and its role in the development of anxiety states. Psychiatr. Dev. 3, 167-226.
Cloninger, C.R., Svrakic, D.M., Przybecky, T.R., 1993. A psychobiological model of temperament and character. Arch. Gen. Psychiatry 50, 975-990.
Corr, P.J., McNaughton, N., 2008. Reinforcement sensitivity theory and personality, in: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality. Cambridge University Press, Cambridge, pp. 155-187.
Covey, D.P., Howard, C.D., 2011. Dopaminergic signaling in cost-benefit analyses: a matter of time, effort, or uncertainty? J. Neurosci. 31, 1561-1562.
Cunningham, W.A., Arbuckle, N.L., Jahn, A., Mowrer, S.M., Abduljalil, A.M., 2010. Aspects of neuroticism and the amygdala: chronic tuning from motivational styles. Neuropsychologia 48, 3399-3404.
Daniel, R., Pollmann, S., 2010. Comparing the neural basis of monetary reward and cognitive feedback during information-integration category learning. J. Neurosci. 30, 47-55.
Darwin, C., 1965/1872. The expression of the emotions in man and animals. University of Chicago Press, Chicago, Illinois.
Davidson, R.J., Ekman, P., Saron, C.D., Senulis, J.A., Friesen, W.V., 1990. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology I. J. Pers. Soc. Psychol. 58, 330-341.
Davidson, R.J., Shackman, A.J., Maxwell, J.S., 2004. Asymmetries in face and brain related to emotion. Trends in Cognitive Sciences 8, 389-391.
De Martino, B., Kumaran, D., Seymour, B., Dolan, R.J., 2006. Frames, biases, and rational decision-making in the human brain. Science 313, 684-687.
De Villiers, P., 1977. Choice in concurrent schedules and a quantitative formulation of the law of effect, in: Honig, W.K., Staddon, J.E.R. (Eds.), Handbook of operant behaviour. Prentice Hall, Englewood Cliffs, pp. 233-278.
Depue, R.A., Collins, P.F., 1999. Neurobiology of the structure of personality: dopamine, facilitation of incentive motivation and extraversion. Behav. Brain Sci. 22, 491-569.
DeYoung, C.G., 2010. Personality neuroscience and the biology of traits. Social and Personality Psychology Compass 4, 1165-1180.

DeYoung, C.G., Gray, J.R., 2009. Personality neuroscience: explaining individual differences in affect, behaviour and cognition, in: Corr, P.J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 323-346.
DeYoung, C.G., Quilty, L.C., Peterson, J.B., 2007. Between facets and domains: ten aspects of the big five. J. Pers. Soc. Psychol. 93, 880-896.
Dickinson, A., 1980. Contemporary animal learning theory. Cambridge University Press, Cambridge.
Duggan, C., Sham, P., Lee, A., Minne, C., Murray, R., 1995. Neuroticism: a vulnerability marker for depression evidence from a family study. J. Affect. Disord. 35, 139-143.
Ferguson, E., Heckman, J., Corr, P.J., 2011. Editorial. Personality and economics: overview and proposed framework. Personality and Individual Differences 51, 201-209.
Finlayson, G., King, N., Blundell, J.E., 2007. Liking vs. wanting food: importance for human appetite control and weight regulation. Neurosci. Biobehav. Rev. 31, 987-1002.
Frey, B.S., Stutzer, A., 2007. Economics and psychology: a promising new cross-disciplinary field. MIT Press, London, England.
Glimcher, P.W., Dorris, M.C., Bayer, H.M., 2005. Physiological utility theory and the neuroeconomics of choice. Games and Economic Behavior 52, 213-256.
Glimcher, P.W., Rustichini, A., 2004. Neuroeconomics: the consilience of brain and decision. Science 306, 447-452.
Graeff, F.G., 1991. Neurotransmitters in the dorsal periaqueductal grey and animal models of panic anxiety, in: Briley, M., File, S.E. (Eds.), New concepts in anxiety. MacMillan Press, pp. 288-312.
Gray, J.A., 1967. Disappointment and drugs in the rat. Adv. Sci. 23, 595-605.
Gray, J.A., 1970. The psychophysiological basis of introversion- extraversion. Behav. Res. Ther. 8, 249-266.
Gray, J.A., 1973. Causal models of personality and how to test them, in: Royce, J.R. (Ed.), Multivariate analysis and psychological theory Academic Press, London, pp. 409-463.
Gray, J.A., 1975. Elements of a two-process theory of learning. Academic Press, London.
Gray, J.A., 1977. Drug effects on fear and frustration: possible limbic site of action of minor tranquilizers, in: Iversen, L.L., Iversen, S.D., Snyder, S.H. (Eds.), Handbook of psychopharmacology.Vol 8. Drugs, neurotransmitters and behaviour. Plenum Press, New York, pp. 433-529.
Gray, J.A., 1981. A critique of Eysenck's theory of personality, in: Eysenck, H.J. (Ed.), A model for personality. Springer, Berlin, pp. 246-276.
Gray, J.A., 1982. The Neuropsychology of Anxiety: an enquiry in to the functions of the septo-hippocampal system. Oxford University Press, Oxford.
Gray, J.A., 1987. The psychology of fear and stress. Cambridge University Press, London.
Gray, J.A., McNaughton, N., 2000. The Neuropsychology of Anxiety: an enquiry into the functions of the septo-hippocampal system. Oxford University Press, Oxford.
Gray, J.A., Smith, P.T., 1969. An arousal-decision model for partial reinforcement and discrimination learning, in: Gilbert, R., Sutherland, N.S. (Eds.), Animal discrimination learning. Academic Press, London, pp. 243-272.

Gray, J.R., Burgess, G.C., Schaefer, A., Yarkoni, T., Larsen, R.J., Braver, T.S., 2005. Affective personality differences in neural processing efficiency confirmed using fMRI. Cognitive Affective and Behavioral Neuroscience 5, 182-190.
Haber, S.N., Calzavara, R., 2009. The cortico-basal ganglia integrative network: the role of the thalamus. Brain Res. Bull. 78, 69-74.
Haegelen, C., Rouaud, T., Darnault, P., Morandi, X., 2009. The subthalamic nucleus is a key-structure of limbic basal ganglia functions. Med. Hypotheses 72, 421-426.
Hall, P.J., Chong, W., McNaughton, N., Corr, P.J., 2011. A neuroeconomic perspective on the reinforcement sensitivity theory of personality. Personality and Individual Differences 51, 242-247.
Hinde, R.A., 1982. Ethology. Fontana.
Hode, Y., Ratomponirina, C., Gobaille, S., Maitre, M., Kopp, C., Misslin, R., 2000. Hypoexpression of benzodiazepine receptors in the amygdala of neophobic BALB/c mice compared to C57BL/6 mice. Pharmacol. Biochem. Behav. 65, 35-38.
Jenison, R.L., Rangel, A., Oya, H., Kawasaki, H., Howard, M.A., 2011. Value encoding in single neurons in the human amygdala during decision making. J. Neurosci. 31, 331-338.
Jocham, G., Klein, T.A., Ullsperger, M., 2011. Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J. Neurosci. 31, 1606-1613.
Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat. Neurosci. 10, 1625-1633.
Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263-291.
Kang, M.J., Rangel, A., Camus, M., Camerer, C.F., 2011. Hypothetical and real choice differentially activate common valuation areas. J. Neurosci. 31, 461-468.
Kapczinski, F., Curran, H.V., Gray, J., Lader, M., 1994. Flumazenil has an anxiolytic effect in simulated stress. Psychopharmacology (Berl). 114, 187-189.
Kendler, K.S., Prescott, C.A., Myers, J., Neale, M.C., 2003. The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. Arch. Gen. Psychiatry 60, 929-937.
Kenny, P.J., 2011. Reward mechanisms in obesity: new insights and future directions. Neuron 69, 664-679.
Killeen, P.R., 1994. Mathematical principles of reinforcement. Behav. Brain Sci. 17, 105-172.
Krueger, R.F., 1999. The structure of common mental disorders. Arch. Gen. Psychiatry 56, 921-926.
Kurth-Nelson, Z., Redish, A.D., 2010. A reinforcement learning model of precommitment in decision making. Frontiers in Behavioral Neuroscience 4, Article 184.
Lehmann, J., Weizman, R., Leschiner, S., Feldon, J., Gavish, M., 2002. Peripheral benzodiazepine receptors reflect trait (early handling) but not state (avoidance learning). Pharmacol. Biochem. Behav. 73, 87-93.
Leue, A., Beauducel, A., 2008. A meta-analysis of reinforcement sensitivity theory: on performance parameters in reinforcement tasks. Personality and Social Psychology Review 12, 353-369.
Loewenstein, G., Scott, R., Cohen, J.D., 2008. Neuroeconomics. Annu. Rev. Psychol. 59, 647-672.

Lykken, D.T., 1971. Multiple factor analysis and personality research. Journal of Research in Personality 5, 161-170.
Mackintosh, N.J., 1974. The psychology of animal learning. Academic Press, New York.
Matthews, G., Gilliland, M.A., 1999. The personality theories of H. J. Eysenck and J. A Gray: a comparative review. Personality and Individual Differences 26, 583-626.
McNaughton, N., 1985. The effects of systemic and intraseptal injections of sodium amylobarbitone on rearing and ambulation in rats. Australian Journal of Psychology 37, 15-27.
McNaughton, N., 2002. Aminergic transmitter systems, in: D'Haenen, H., Den Boer, J.A., Westenberg, H., Willner, P. (Eds.), Textbook of Biological Psychiatry. John Wiley & Sons, pp. 895-914.
McNaughton, N., Corr, P.J., 2004. A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance. Neurosci. Biobehav. Rev. 28, 285-305.
McNaughton, N., Corr, P.J., 2008a. Animal cognition and human personality, in: Corr, P.J. (Ed.), The Reinforcement Theory of Personality. Cambridge University Press, Cambridge, pp. 95-119.
McNaughton, N., Corr, P.J., 2008b. RST and personality, in: Corr, P.J. (Ed.), The Reinforcement Theory of Personality. Cambridge University Press, Cambridge.
McNaughton, N., Gray, J.A., 1983. Pavlovian counterconditioning is unchanged by chlordiazepoxide or by septal lesions. Q. J. Exp. Psychol. 35B, 221-233.
McNaughton, N., Kocsis, B., Hajós, M., 2007. Elicited hippocampal theta rhythm: a screen for anxiolytic and pro-cognitive drugs through changes in hippocampal function? . Behav. Pharmacol. 18, 329-346.
McNaughton, N., Owen, S., Boarder, M.R., Gray, J.A., Fillenz, M., 1984. Responses to novelty in rats with lesions of the dorsal noradrenergic bundle. New Zealand Journal of Psychology 13, 16-24.
Meyer, T.J., Miller, M.L., Metzger, R.L., Borkovec, T.D., 1990. Development and validation of the Penn state worry questionnaire. Behav. Res. Ther. 28, 487-495.
Millenson, J.R., Leslie, J.C., 1979. Principles of behavior analysis. MacMillian, New York.
Miller, N.E., 1944. Experimental studies of conflict, in: Hunt, J.M. (Ed.), Personality and the behavioural disorders. Ronald Press, New York, pp. 431-465.
Miller, N.E., 1959. Liberalization of basic S-R concepts: extensions to conflict behaviour, motivation and social learning, in: Koch, S. (Ed.), Psychology: a study of a science. Wiley, New York, pp. 196-292.
Mitchell, D.J., McNaughton, N., Flanagan, D., Kirk, I.J., 2008. Frontal midline theta from the perspective of hippocampal "theta". Prog. Neurobiol. 86, 156-185.
Monosov, I.E., Hikosaka, O., 2012. Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex. J. Neurosci. 32, 10318-10330.
Montagna, P., Sforza, E., Tinuper, P., Provini, F., Plazzi, G., Cortelli, P., Schoch, P., Rothstein, J.D., Lugaresi, E., 1995. Plasma endogenous benzodiazepine-like activity in sleep disorders with excessive daytime sleepiness. Neurology 45, 1783-1783.
Moore, R.A., Gale, A., Morris, P.H., Forrester, D., 2006. Theta phase locking across the neocortex reflects cortico-hippocampal recursive communication during goal conflict resolution. Int. J. Psychophysiol. 260, 260–273.

Moore, R.A., Mills, M., Marsham, P., Corr, P.J., 2012. Behavioural Inhibition System (BIS) sensitivity differentiates EEG theta responses during goal conflict in a continuous monitoring task. Int. J. Psychophysiol. 85, 135-144.
Neo, P.S.H., 2008. Theta activations associated with goal-conflict processing: evidence for the revised Behavioural Inhibition System., PhD Thesis, Department of Psychology. University of Otago, Dunedin.
Neo, P.S.H., Thurlow, J., McNaughton, N., 2011. Stopping, goal-conflict, trait anxiety and frontal rhythmic power in the stop-signal task. Cognitive Affective & Behavioral Neuroscience 11, 485-493.
Novemsky, N., Kahneman, D., 2005. The boundaries of loss aversion. Journal of Marketing Research 42, 119–128.
Okaichi, Y., Okaichi, H., 1994. Effects of fimbria-fornix lesions on avoidance tasks with temporal elements in rats. Physiol. Behav. 56, 759-765.
Ostlund, S.B., Wassum, K.M., Murphy, N.P., Balleine, B.W., Maidment, N.T., 2011. Extracellular dopamine levels in striatal subregions track shifts in motivation and response cost during instrumental conditioning. J. Neurosci. 31, 200-207.
Padoa-Schioppa, C., 2011. Neurobiology of economic choice: a good-based model. Annu. Rev. Neurosci. 34, 333-359.
Perkins, A.M., Ettinger, U., Davis, R., Foster, R., Williams, S.C.R., Corr, P.J., 2009. Effects of lorazepam and citalopram on human defensive reactions: ethopharmacological differentiation of fear and anxiety. J. Neurosci. 29, 12617-12624.
Perkins, A.M., Ettinger, U., Williams, S.C.R., Reuter, M., Hennig, J., Corr, P.J., 2011. Flight behaviour in humans is intensified by a candidate genetic risk factor for panic disorder: evidence from a translational model of fear and anxiety. Mol. Psychiatry 16, 242-244.
Pickering, A.D., Corr, P.J., Powell, J.H., Kumari, V., Thornton, J.C., Gray, J.A., 1997. Individual differences in reactions to reinforcing stimuli are neither black nor white: to what extent are they gray?, in: Nyborg, H. (Ed.), The Scientific Study of Human Nature: Tribute to Hans J. Eysenck at Eighty. Elsevier, London, pp. 36-67.
Polc, P., 1995. Involvement of endogenous benzodiazepine receptor ligands in brain disorders: therapeutic potential for benzodiazepine antagonists. Med. Hypotheses 44, 439-446.
Powell, G.E., 1979. Brain and personality. Saxon House, London.
Redgrave, P., Gurney, K., 2006. The short-latency dopamine signal: a role in discovering novel actions? Nat Rev Neurosci 7, 967-975.
Reuter, M., 2008. Neuro-imaging and genetics, in: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality. Cambridge University Press, Cambridge, pp. 317-344.
Rovner, B.W., Casten, R.J., 2001. Neuroticism predicts depression and disability in age-related macular degeneration. J. Am. Geriatr. Soc. 49, 1097-1100.
Roy, A., 1999. Neuroticism and depression in alcoholics. J. Affect. Disord. 52, 243-245.
Rushworth, M.F.S., Behrens, T.E.J., 2008. Choice, uncertainty and value in prefrontal and cingulate cortex. Nature Neurosci. 11, 389-397.
Sallet, J., Rushworth, M.F.S., 2009. Should I stay or should I go: genetic bases for uncertainty-driven exploration. Nature Neurosci. 12, 963-965.
Sanfey, A.G., Loewenstein, G., McClure, S.M., Cohen, J.D., 2006. Neuroeconomics: cross-currents in research on decision-making. Trends in Cognitive Sciences 10, 108-116.

Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. Annu. Rev. Psychol. 57, 87-115.
Smillie, L.D., 2008. What is reinforcement sensitivity? Neuroscience paradigms for approach-avoidance process theories of personality. European Journal of Personality 22, 359-384.
Smillie, L.D., Jackson, C.J., Dalgleish, L.I., 2006a. Conceptual distinctions between Carver and White's (1994) BAS scales: a reward-reactivity versus trait impulsivity perspective. Personality and Individual Differences 40, 1039-1050.
Smillie, L.D., Pickering, A.D., Jackson, C.J., 2006b. The new reinforcement sensitivity theory: implications for personality measurement. Personality and Social Psychology Review 10, 320-325.
Sokol-Hessner, P., Hsu, M., Curley, N.G., Delgado, M.R., Camerer, C.F., Phelps, E.A., 2009. Thinking like a trader selectively reduces individuals' loss aversion. Proc. Natl. Acad. Sci. U. S. A. 106, 5035-5040.
Spielberger, C.D., Gorusch, R.L., Lushene, R., Vagg, P.R., Jacobs, G.A., 1983. Manual for the STATE-TRAIT ANXIETY INVENTORY (Form Y). Consulting Psychologists Press, Palo Alto, CA94306.
Sylvers, P., Lilienfield, S.O., LaPrairie, J.L., 2011. Differences between trait fear and trait anxiety: implications for psychopathology. Clin. Psychol. Rev. 31, 122-137.
Thibodeau, R., 2010. Approach and withdrawal actions modulate the startle reflex independent of affective valence and muscular effort. Psychophysiology 48, 1011-1014.
Tom, S.M., Fox, C.R., Trepel, C., Poldrack, R.A., 2007. The neural basis of loss aversion in decision-making under risk. Science 315, 515-517.
Torrubia, R., Avila, C., Caseras, X., 2008. Reinforcement sensitivity scales, in: Corr, P.J. (Ed.), The Reinforcement Sensitivity Theory of Personality Cambridge University Press, Cambridge, pp. 188-226.
Torrubia, R., Ávila, C., Moltó, J., Caseras, X., 2001. The sensitivity to punishment and sensitivity reward questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. Personality and Individual Differences 31, 837-862.
Treadway, M.T., Zald, D.H., 2011. Reconsidering anhedonia in depression: lessons from translational neuroscience. Neurosci. Biobehav. Rev. 35, 537-555.
Trepel, C., Fox, C.R., Poldrack, R.A., 2005. Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. Cognitive Brain Research 23, 34-50.
Tversky, A., Kahneman, D., 1991. Loss aversion in riskless choice: a reference dependent model. The Quarterly Journal of Economics 106, 1039-1061.
Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297–323.
Vickery, Timothy J., Chun, Marvin M., Lee, D., 2011. Ubiquity and Specificity of Reinforcement Signals throughout the Human Brain. Neuron 72, 166-177.
Wacker, J., Mueller, E.M., Hennig, J., Stemmler, G., 2012. How to consistently link extraversion and intelligence to the catechol-o-methyltransferase (COMT) gene: on defining and measuring psychological phenotypes in neurogenetic research. J. Pers. Soc. Psychol. Advance online publication.
Wang, Z., Valdes, J., Noyes, R., Zoega, T., Crowe, R.R., 1998. Possible association of a cholecystokinin promotor polymorphism (CCK-36CT) with panic disorder. Am. J. Med. Genet. 81, 228-234.
Weller, J.A., Levin, I.P., Shiv, B., Bechara, A., 2007. Neural correlates of adaptive decision making for risky gains and losses. Psychological Science 18, 958-964.

Young, C.K., McNaughton, N., 2009. Coupling of theta oscillations between anterior and posterior midline cortex and with the hippocampus in freely behaving rats. Cereb. Cortex 19, 24-40.
Zak, P.J., 2004. Neuroeconomics. Philosophical Transactions of the Royal Society of London: B 359, 1737-1748.
Zald, D., Depue, R., 2001. Serotonergic modulation of positive and negative affect in psychiatrically healthy males. Personality and Individual Differences 30, 71-86.
Zeiler, M., 1977. Schedules of reinforcement: the controlling variables, in: Honig, W., Staddon, J.E.R. (Eds.), Handbook of operant behavior. Prentice-Hall, Inc., Englewood Cliffs, NJ, pp. 201-232.

FIGURE LEGENDS

Figure 1. The relationship of external items to their internal value, which controls the strength (but not direction) of their effect on behaviour. Individual items (consumable food, dollar gains, etc) have a value that depends on their amount and the current level of specific drive for that class of item. The amount, therefore, interacts with an individual exchange rate (first grey rectangle, represented by varying arrow thicknesses) to generate an internal valuation, which can be positive or negative depending on the valence of the item (e.g. dollar gain versus dollar loss in the form of explicit removal from an existing store). For the same amount of the same class of item, such as dollars (which necessarily matches individual item exchange rate), negative valence has a higher exchange rate (note thicker arrow) and so generates a greater internal value than positive (second grey rectangle, e.g. loss aversion). Loss aversion is a relative term (comparing the effect of loss with that of the same external value of gain) and we take it to represent the difference between trait gain sensitivity and trait loss sensitivity, with the latter being the greater.

Figure 2. The combination of valuation and operant factors that determines response strength and direction. Items with a specific external value ($1) that can be gained or lost are represented by a particular internal amount that will depend on the exchange rate (or the level of “hunger”) for the item (see Figure 1). In this example all inputs are $1 and so exchange rate is ignored. The internal value that drives decisions and the intensity of action also depends on whether the item is gained or lost. Economic analysis has shown that the same external value generally has a greater effect if it is a loss (×Loss exchange rate) than if it is a gain (×Gain exchange rate). The effect of this internal valuation on behaviour then depends on the consequences of responding. Gain production and loss prevention activate approach; loss production and gain prevention activate avoidance. Concurrent APPROACH and AVOID tendencies are then integrated to determine the direction and strength of responding. A fixed internal value of approach and avoidance will have different effects on response strength (×Attractor exchange rate; × Repulsor exchange rate) that depends both on factors of reinforcement sensitivity and on the distance from the goal that will be achieved by responding. (Approach and avoidance have different goal gradients, see Section 2.8.)

Figure 3. A. Observed speeds resulting from the combination of a specific dollar value (gain averaged with omission of loss – separate values plotted as separate curves) for the production of a response with a specific dollar value (loss averaged with the omission of gain – repulsor value, X axis) for the inhibition of the response. The probability for gain or loss on any particular trial was equal. Open circles indicate the point on each curve at which the net value averages to $1. B. The same data represented as point values with the curves resulting from the optimised fitted functions based on previous animal behaviour analysis (see text).

Figure 4. Observed differences in speed between gain and loss manipulations for the combination of a specific attraction value for the response with a particular repulsion value. Attraction dollar value is gain averaged with omission of loss (separate values plotted as separate curves). Repulsion dollar value is loss averaged with the omission of gain and is plotted on the X axis. The probability of gain or loss on any particular trial was equal. Open circles indicate the point on each curve at which the attraction and repulsion values are equal.

Figure 5. Overall relation of the BIS, FFFS and BAS – an updated model. To activate the BIS one must generate concurrent and approximately equal activation of the FFFS and the BAS, i.e. face the animal with an approach-avoidance conflict. Both simple approach and simple avoidance will then be inhibited and replaced with environmental scanning (in the form of altered attention), external scanning (risk assessment behaviour) and internal scanning of memory. Note that all of these scanning operations are aimed at detecting affectively negative information and involve a selective increase (stippled arrow) in the salience and value of aversive information. As a result, a secondary consequence of activation of the system is normally a shift of the balance between approach and avoidance tendencies in the direction of avoidance. However, when scanning determines that danger is absent the approach-avoidance conflict is resolved in favour of approach. The inputs to the system are classified in terms of the delivery (+) or omission (-) of primary positive reinforcers (PosR) or primary negative reinforcers (NegR) or conditional stimuli (CS) or innate stimuli (IS) that predict such primary events. (Adapted from Gray and McNaughton, 2000.) As discussed earlier we see loss (i.e. removal of a positive reinforcer from

an existing store) as a form of NegR, thus allowing for both Loss+ and Loss- (see section 2.6, 2.7). Specific cases of PosR and NegR will have their own individual exchange rates but, as discussed in the text, their effect will also be modulated by a general sensitivity factor that is different for the two classes of reinforcer. The stippled areas in the model are all points at which general personality factors could operate (see section 4).