Neuroscience and Biobehavioral Reviews 191 (2026) 106926

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

journal homepage: www.elsevier.com/locate/neubiorev

Psychological traits and their explanation: A 20-year perspective from Reinforcement Sensitivity Theory

Philip J. Corr [a,c], Neil McNaughton [b,*]

[a] Department of Psychology and Neuroscience, City St Georges, University of London, UK
[b] Department of Psychology, University of Otago, Dunedin, New Zealand
[c] Walbrook Institute London, UK

ARTICLE INFO

Keywords:
Reinforcement sensitivity theory
Personality explanation
Neurology of traits

ABSTRACT

Reinforcement Sensitivity Theory (RST) attributes certain personality traits to sensitivities of neural systems based on a detailed long-lasting neuropsychology of state control. (First book-length edition, 1982, 3rd edition, 2024). However, for trait system sensitivities, within-system details can largely be ignored. Here we provide an overview of the systems for personality researchers. We also extended the underdeveloped personality aspects of the neuropsychology. In the process of producing this extension, we discovered various lacunae in personality theory as a whole (and the role of scientific explanation within it) that has led us to start the paper with a detailed exposition of application of levels of explanation to personality, the role of emergence in such explanations, and (we believe) a resolution among the different schools of personality research (lexical, rational, neural). [128/250]

1. Introduction

‘When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘It means just what I choose it to mean – neither more nor less.’

‘The question is,’ said Alice, ‘whether you can make words mean so many different things.’

‘The question is,’ said Humpty Dumpty, ‘which is to be Master - that's all.’

Lewis Carroll: Through the Looking Glass

Reinforcement Sensitivity Theory (RST) attributes certain personality traits to sensitivities of neural systems. The system architectures are based on a detailed neuropsychology of state control that has a recent book-length update (McNaughton and Gray, 2024). However, with a focus on personality, these within-system details can largely be ignored as the sensitivity of a system as a whole is what matters.

Here, we explain how the sensitivity of these systems to their specific inputs can account for the development and expression of specific related traits. However, to do this, we discovered that we first needed to consider traits in general and psychological, specifically personality, ones in particular. We also needed to look at how such high-level personality traits could, in principle, be explained by the general sensitivities of lower-level neural systems. To take this beyond the general, we provide a specific explanation of the psychological trait of goal inhibition sensitivity – the sensitivity of the Goal Inhibition System (McNaughton and Gray, 2024), originally named the Behavioural Inhibition System (Gray, 1969, 1970b, 1982; Gray and McNaughton, 2000). We believe that starting with a well-developed neurally-detailed state system provides the detailed machinery that has the unique capacity to explain a limited set of motivation-related traits at the neural level. We hope that this specific example can then provide a general model for explanation of other traits.

So, what is a trait? With trepidation, we will attempt to bring some order to a field rife with ambiguity and semantic confusion. Is “a trait” just the observation of a superficial character such as eye colour? More generally, to talk about a trait, in contrast to a feature, is to imply, at the least, regularities in individual or species variation. Such regularities imply some controlling factor that provides an explanation for them and the clustering of features that results. Or, else, is a trait just a set of

* Correspondence to: Department of Psychology, University of Otago, PO Box 56, Dunedin 9054, New Zealand.
E-mail address: neil.mcnaughton@otago.ac.nz (N. McNaughton).

https://doi.org/10.1016/j.neubiorev.2026.106926

Available online 25 August 2026

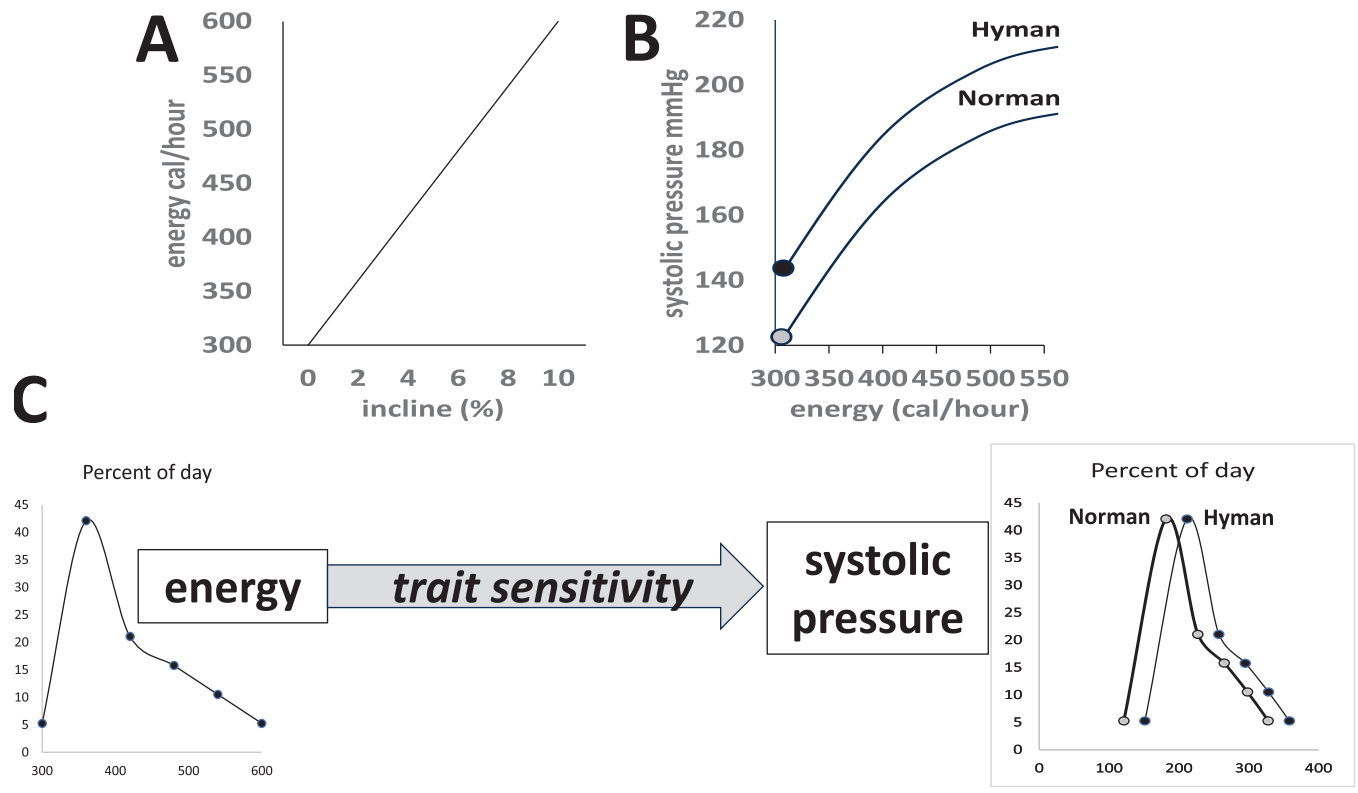

Fig. 1. Trait sensitivity. External circumstance (treadmill incline/ energy) determines state output (systolic pressure) that depends on an individual's (trait) sensitivity. Strictly, this sensitivity is a rate parameter. But clinically, trait blood pressure is an average of resting state measures (dots). Note that state variation vastly exceeds individual resting differences. Nominal males are shown; female values would likely be lower.

externally observable superficial regularities, such as seen in personality factors, or is it the endogenous cause of them? These descriptive and causal issues are problematic. Description and cause often get conflated or confused with each other when we talk of an 'explanation' of personality.

There is a further conundrum: Is what is meant by "a trait" in psychology different from "a trait" in the rest of biology? Whereas biology usually starts with simple single-feature physical measures, psychology (whether via behaviour or questionnaires) must deal with more complex, but consistent, patterns of behavioural or mental features. The regularity of such general patterns can allow some prediction via generalisation. But, as with the biological, is the trait the observed pattern (the easiest to measure) or its endogenous cause (more difficult to discern)? And, to what does the name of a specific "psychological trait" refer, let alone explain? Conversely, how best can we characterise and explain a psychological trait, and what need to be considered different types? Indeed, should we use several different definitions of "trait" at the same time – and if so, how do we keep track of causal and descriptive aspects? These simple questions do not invite simple answers.

So, to ask these basic questions is to enter a fantastical land, where trying to run fast leaves you in the same place. Lewis Caroll is our guide. Our goal, like Humpty Dumpty, is to be Master of the words we use. But circumstances forced us to start with *Jabberwocky*; and, it turns out, finish with *Alice's Adventures in Wonderland*, where "Everybody has won, and all must have prizes".

Much of what follows relates to various aspects of RST; but this has evolved through successive incarnations, so the next section briefly summaries this evolution. Also, to help us gain traction in this slippery world, we will take a step backwards into basic biology before moving forward to ask psychological questions. As we do so, we should remember that many physical traits (or their endophenotypes) can be measured directly; in contrast to most psychological ones. That said, some physical trait constructs require inference even when their states can be describe directly (we will use blood pressure as an example). This shows how state and trait can often be conflated and confused. Such physical biological traits show us clearly that this is a problem. But, because they are observationally more complex, it is much more fraught with psychological ones. Specifically, the extra complication with psychological traits is that the states to which they contribute must be assessed indirectly, using statistics that condense multiple direct behavioural measures (whether verbal or non-verbal). The situation would not be out of place in Wonderland.

2. Recap of the RST of personality

Jeffrey Gray (1960), (1970b) took an explicitly neural (and pharmacologically anchored) approach to state systems. However, trait RST derives from his foundational paper (Gray, 1970a) that provided no neural explanation for the system sensitivities invoked to explain the psychological nature of the traits. This left open the question of what is meant, in neural terms, by RST from the standpoint of a personality theorist. For historical reasons, there has been confusion here.

The detailed *state neuropsychology* (Gray, 1982) was updated in a 2nd edition nearly 20 years later, with considerable theory development (Gray and McNaughton, 2000) and with a further 25 years of theory development in a 3rd edition (McNaughton and Gray, 2024). Throughout this development, the core of the theory was three motivational systems – but with new data forcing addition of neural components and changes in their nominal organisation and their detailed psychological functions.

In the most recent edition, the core of the neuropsychology is a Goal Inhibition System, (GIS; processing goal conflict and passive avoidance, previously named the Behavioural Inhibition System, BIS) with an associated Goal Attraction System (GAS, processing approach; previously named the Behavioural Approach System, BAS) and Goal

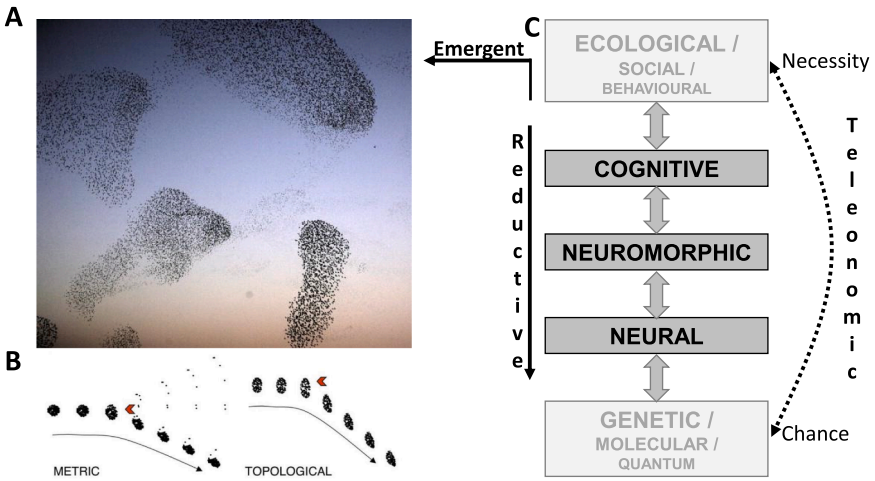

Fig. 2. A. Murmuration of starlings.There is no direct control of the shape of a murmuration it emerges from the mass of small-scale individual starling interactions. Photo: courtesy of ISC-CNR, Starflag Project from King and Sumpter (2012) via Elsevier user licence. B. Simulations comparing metric and topological individual interactions. Starlings follow a topological rule, interacting with a fixed 6–7 adjacent birds – the simulation shows this make the flock cohesive when attacked by a hawk (◀). From Ballerini et al. (2008) via PNAS Open policy. C. Levels of explanation for personality neuroscience. Conventional reductive explanation for phenomena at the observational (including social and ecological) levels can usefully descend through cognitive and neuromorphic levels to constructs at the neural level (McNaughton and Smillie, 2018). That is, a detailed theory expressed in neural terms can, potentially, fully explain an observation at the personality level. However, at each of these levels, emergent properties may need to be invoked, and with some aspects of behaviour, a full explanation may include emergent properties at the behavioural level. In general, social and ecological factors will operate at too high a level to provide useful detailed explanation of personality while genetics, molecular (biochemistry and chemistry) and quantum mechanics will usually provide details that confuse more than they explain. Teleonomic (evolutionary) explanations span and link the other levels of explanation. Figure and legend from McNaughton (2020) under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/).

Repulsion System (GRS, processing escape and active avoidance; previously named for its lowest levels: the Flight-Flight-Freeze system, FFFS). The change in names reflects a move away from systems defined by rat behavioural neuropsychology – to ones emphasizing internal motivational states (with goals resulting in internal motivational gradients). The change in system names described above was driven by two facts. First, some examples of what one might want to call *behavioural* inhibition (particularly stopping in the stop signal task) proved insensitive to the anxiolytic drugs that define the GIS/BIS. Second the most commonly used "BIS" questionnaire scale does not measure the neurally-defined system (see below). In general, personality theory has tended to lag behind developments in the state neuropsychology – with the state theory of the 2nd edition still not fully transferred to "RST", as usually interpreted by personality theorists.

Twenty years ago, we published summaries of the then-current state of play of RST and of the neuropsychology and anxiety (Corr, 2004; McNaughton and Corr, 2004). Here we provide the 20-year update – more closely integrating the neuropsychological and personality components. For this update, given the current problems with integration, we found we needed not just to link the state theory of the 3rd edition to trait RST but also to fill in, as far as possible, the neural gaps in the trait theory. That is, we try and indicate how traits are instantiated within the state neurology. How are traits, within current versions of RST, anchored? This specific, RST-driven, attempt has forced us to also revisit what is meant by a trait in the context of personality theory more generally. A key question is, can we have a trait theory (describing individual differences) without a theory of the state variations that supply the relevant individual differences?

A crucial point here, is that personality factors cannot *reduce* to the state systems for which they provide the control of sensitivity. For a personality factor to be a system *sensitivity* it, itself, must be outside the system as a whole. But the two must overlap. This is most obvious with the GIS. Here the system is *defined* by the action of anxiolytic drugs. By the 3rd edition, these had forced inclusion in the GIS of a wide range of parts (Fig. 6). Anxiolytics act *on all* of these interconnected areas. So, the neural system (which encompasses multiple components) can be defined both by its overall function of resolving goal conflict and by its general sensitivity to what we can presume will be endogenous benzodiazepine receptor ligands in the blood stream. One obvious source of trait sensitivity for RST systems, then, will be hormones. However, it should be noted (see section on goal inhibition below) that diffuse neural systems controlling, e.g., rhythmicity could be equally effective. Critically, the ascending monoamine systems can be viewed as neural methods of delivering hormones more locally than does the blood stream (via varicosities that are less local than synapses). Whether delivered via nerves or blood stream, the trait control is chemical and massively distributed. These chemicals then alter the global sensitivity of extensively interconnected systems, determining which of their parts will generate behavioural output given any particular eliciting condition.

So, to bring traits of personality and states into closer register with RST, we have two tasks to complete. The first is to present the salient aspects of the revised version of RST (McNaughton and Gray, 2024); and, second, to relate this revised neuropsychology to human personality traits. Partly, these tasks reflect the traditional dichotomy between 'neuroscience' and 'personality' (Corr and Mobbs, 2018). It is sobering to reflect that, despite the many years that have passed since the inception of the first book-length formulation of RST (Gray, 1982), and the many hundreds of papers using this 'personality theory', we are still left to ask basic questions about its trait representation. Be that as it may, our task is relevant because our specific concerns have implications for the general question of how biological traits relate to psychological ones. It is also becoming clear that the biology underlying psychological traits is likely to be crucial to our capacity to treat mental (as opposed to neurological) disorders – with disorders such as PTSD reflecting major changes in personality.

3. What do you mean by "a trait" in biology?

Given a neurobiological starting point, "a trait" is a relatively straightforward concept but still needs some unpacking. For the neurobiologist, it is

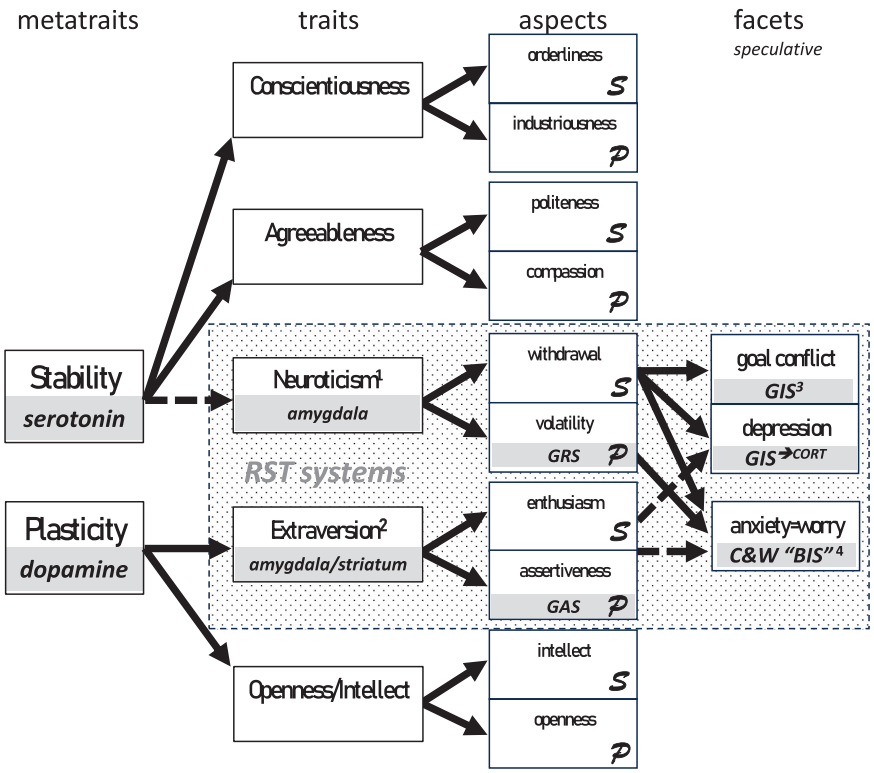

Fig. 3. Cybernetic Big 5 correlations (dashed lines are negative) with some neural attributions. The main metatrait, trait, and aspect content of this figure is derived from DeYoung (2015). Stability/plasticity can be seen as synonymous with resilience/efficiency as fundamental evolutionary constraints (Hovhannisyan and Vervaeke, 2021). DeYoung's cybernetic version of the Big 5 focuses on goal control for traits (as does RST for states). There are detailed known neural correlates of all the Big 5 metatraits and traits (Allen and DeYoung, 2017; DeYoung et al., 2021). Our grey box neural attributions to metatraits and traits focus only on those that map to sensitivities of state RST-related structures. We have speculatively added some RST-related aspect links and a few possible facets with links based lexically on aspect descriptions in DeYoung (2015), Table 1, p. 42) (reformatted in Allen and DeYoung, 2017). All neural annotations reflect system sensitivities with serotonin and dopamine controlling a wide range of these although correlation is only shown at the trait level (*P* = additional aspect level influence of plasticity; S = additional aspect level influence of stability). Notes: 1) Increased amygdala sensitivity is matched by decreased control from DL/VL PFC; this is sensitivity of the extended amygdala (including the basal nucleus of the stria terminalis) with the sensitivities of all the nuclei involved presumed changed by ketamine. 2) Orbital frontal cortex contributes to higher order, e.g., social components. 3) As defined by anxiolytic drugs. 4) This does NOT relate to the GIS. Abbreviations: C&W "BIS" = trait measured by C&W scale; GAS = Goal Approach System sensitivity; GIS = Goal Inhibition System sensitivity; GRS – Goal Repulsion System sensitivity; ➨^{CORT} = release of CORT.

> "…a specific characteristic of an individual. Traits can be determined by genes, environmental factors or by a combination of both. Traits can be qualitative (such as eye color) or quantitative (such as height or blood pressure). A given trait is part of an individual's overall phenotype." [https://www.genome.gov/genetics-glossary/Trait].

From this definition, it is clear that such traits are measurable. Many easily. Their causation is usually much more obscure. But note the emphasis of the quote on causality. The quote is clear that traits are determined biologically via inheritance and development. Together, traits combine to comprise the phenotype.

The quoted definition treats eye colour as qualitative (blue, green, brown, and several more) but this categorisation is due to how colour is perceived by humans, not to the endogenous trait giving rise to the colour: melanin concentration. This varies from low (blue eyes) through medium (green eyes) to high (brown eyes). Once we have identified this underlying cause of apparently categorical colour, we can (as with psychological scale scores) assign a number to the strength of the underlying trait. Similar category versus dimension perceptual issues will occur with human personality measures, particularly lexically-derived ones that reflect a subjective human interpretation of patterns in a proximal environment. As with melanin concentration, the underlying causation might be very different to what is perceived by the observer.

With some quantitative traits like height, we can ask two sorts of causal question: (1) What causes the individual's trait value (in terms of measured numerical value?); and (2) What negative effects do high (or low) levels of the trait cause? An example of a question about *non-genetic* causes of traits would include: What causes reduced height relative to parents (with dietary answers being particularly interesting)? Key questions about effects of high (or low) traits would include: What problems do excess height cause (answers include difficulty negotiating doorways, headache, and joint pain)? It is the combination of the answers to 1 and to the "high" and "low" versions of 2 that account, in an evolutionary context, for the distribution of values of the trait across individuals in a population. In relation to the second question, this takes us into the personality and psychopathology domain.

Thus far, we have said little more than that we can measure biologically important parameters, often themselves obliquely related (e.g., height and weight). But when we name these parameters 'traits', we usually imply that the differences in trait value between individuals tells us something important about things like their medical status and likely life outcomes – they are not merely descriptive without inferential value. For example, apparent eye colour tells us very little of causal relevance; melanin concentration much more so, especially if we discover it is related to other important biological processes and traits (e.g., related to UV-sensitivity of the skin). As we will see, the same appears true of psychological traits.

4. Trait versus state in biology

Note that the traits listed above are all linked directly to measurable

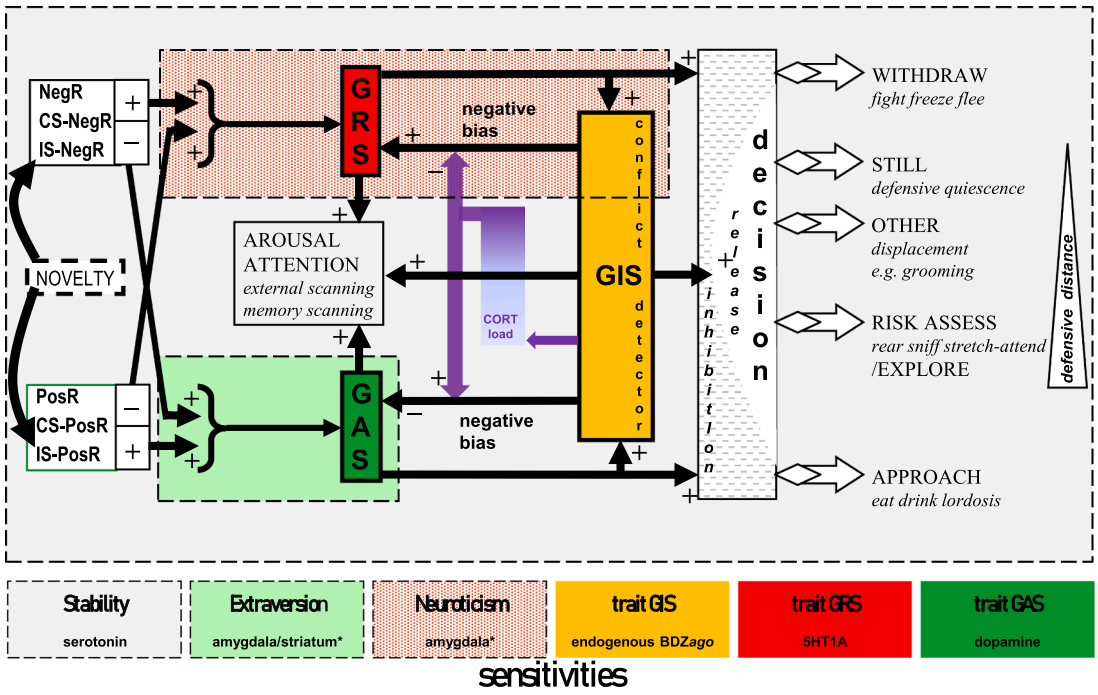

Fig. 4. Relationship of relevant system sensitivities (traits) to the state control of the Goal Attraction System (GAS), Goal Repulsion System (GRS) and Goal Inhibition System (GIS). The key point for state control is that the GIS is activated when the GRS and GAS are in conflict. Resultant behaviours depend on defensive distance, see text. This depends on the sensitivities (colour coded) of the relevant systems. Extensively adapted from McNaughton and Gray (2024) with permission via author's contract. 5HT1A = serotonin (5HT) 1A receptor; + = presentation; - = omission; BDZago = benzodiazepine receptor agonist; CS = conditioned stimulus; GAS = Goal Attraction System; GIS = Goal Inhibition System; GRS = Goal Repulsion System; IS = innate stimulus; NegR = motivationally negative reinforcer; PosR = motivationally positive reinforcer;.

physical characters (wavelength of light, linear distance, or mmHg). Eye colour and, once growth is complete, height have essentially unchanging values. While height does change over the course of the day, for medical purposes this can be treated as measurement error. So, for these measures, there is often no need to distinguish trait from state: to know one is to know the other, at least in their conventional form of expression. However, blood pressure, for example, is different. Unlike height, blood pressure can vary markedly on a timescale of milliseconds. This state variation is a result of clear imposed moment-to-moment circumstances such as an alcoholic drink or a physical load in a treadmill test. In contrast, trait blood pressure, as part of the overall phenotype, can be thought of as a fixed value linked to something like average blood pressure over a range of circumstances.

So, to say someone "has high blood pressure", we can mean either of two things. For someone on a treadmill, we can mean that their blood pressure, right now, is high relative to resting value – this is state blood pressure. For someone being tested by a doctor after a sufficient period of relaxation, we will mean not just that their resting blood pressure is high relative to others in the population but also that their state blood pressure at other times (particularly under stress) will be relatively higher than others. Note that resting blood pressure, in this context, is a convenient indirect clinical means of assessing this heightened system *sensitivity* – this is trait blood pressure and is essentially an inferred construct mainly dependent on the hormone aldosterone (which controls blood volume).

Fig. 1A shows how increasing the incline of a treadmill increases the *energy* required to walk on it by the same amount for all people of a fixed body weight. The energy load interacts with heart control systems to produce a non-linear increase in state systolic blood pressure (Fig. 1B) that *differs* between individuals. Normal trait sensitivity (Norman) produces lower blood pressures relative to high trait sensitivity (Hyman). The circumstances that generate the energy load in real life are multifarious and vary across the day. But they would produce different pressure distributions for Norman and Hyman (Fig. 1C). It is the consistency of the *pattern* of general state function combined with the differences in this pattern between individuals that allows us to infer individual trait control.

But seeing the pattern through the hustle and bustle of everyday life is hard. Treadmills are used to reduce, via standardisation, this "input uncertainty" problem. This idea is particularly important when the measured outputs that are affected by a trait (in contrast to varying state blood pressure) are multiple. Thus, trait sensitivity sits at a central control point in a spider-web like connection of input environmental causal parameters with output response measures.

In medical and physiotherapy practice, whether "blood pressure" means trait or state depends on when and how it is being measured. To estimate week-to-week variations in trait blood pressure (particularly after starting a new medication), the patient needs to remain quiet and seated for at least 5 min before measuring "state" blood pressure at least twice. In measurement terms, testing at rest and taking multiple measurements and averaging removes the moment-to-moment state effects of challenging loads. Note that the trait value does not relate in any simple way to the average state value over time: it is easy enough to reduce the state average by avoiding intense exercise (top end of pressure curves in Fig. 1C). This would not affect the true trait value (linked to aldosterone), but it would confuse the descriptive measurement of it – hence the need for data such as Fig. 1B.

Of course, a patient with high trait blood pressure should want to reduce it. It causes a range of cardio-vascular problems and can produce strokes. Reducing the trait value is hard, though. It needs a mass of lifestyle changes (such as reducing weight and avoiding alcohol) that are usually sufficiently onerous for many to prefer medication. That said, we can account for the presence of a trait distribution in the population because low trait blood pressure has its own problems: dizziness, fainting, heart palpitations, etc. Similar considerations apply to, for example, goal inhibition sensitivity, worry, and related defences. Too high a sensitivity is generally incapacitating – preventing even adaptive action. Too low a sensitivity can be lethal (e.g., when driving).

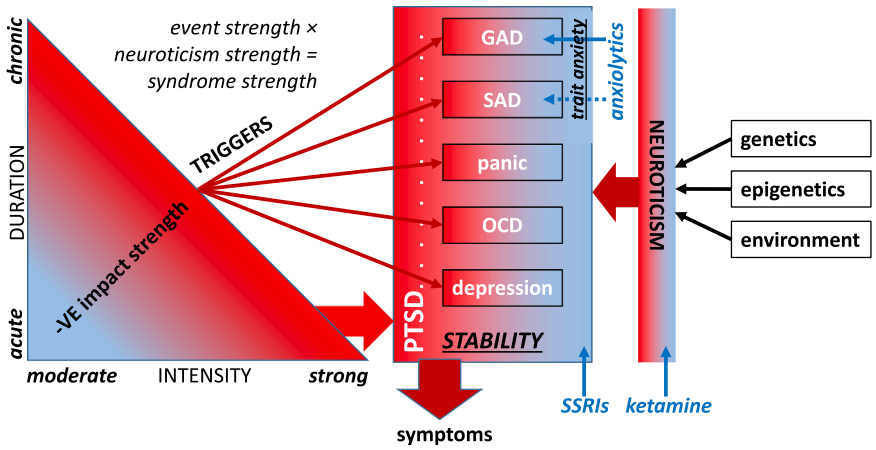

Fig. 5. Diagrammatic representation of the 2-hit hypothesis for disorders known to be affected by ketamine. Disorder is held to result when high neuroticism (resulting from genetic, epigenetic, and prior environmental factors) is combined with a high level of a specific trait linked to Generalized Anxiety Disorder (GAD), Social Anxiety Disorder (SAD), panic, Obsessive Compulsive Disorder (OCD), or depression. High levels of these specific traits can be triggered by moderate chronic or strong acute stress. In the latter case Post-Traumatic Stress Disorder (PTSD) may result. Specific anxiolytic drugs affect the specific trait linked to GAD and to a lesser extent SAD. Specific Serotonin Reuptake Inhibitors (SSRIs) act on a background higher order trait of "stability" (DeYoung, 2006), which would not be specific to aversive events (Carver et al., 2008) and which would slowly alter the specific trait levels. Ketamine acts to affect neuroticism and so alters the disordered expression of all the specific traits. Figure and adapted legend from McNaughton and Glue (2020) with permission via http://creativecommons.org/licenses/by/4.0/.

Unfortunately, like physical traits, mental traits are hard to change.

5. What do we mean by explanation?

If the definition and characterisation of a trait is hard to conceptualise, so too is what is meant by explanation. We turn to this now. "Explanation" is used in many different, sometimes opposing, ways (evolutionary function; immediate cause; detailed mechanism). We take it as given that, if we are doing science rather than bean counting, we are required to produce a predictive explanation at some point. To anticipate what is to follow, we see a major advantage of RST as its provision of machinery that can provide a causal explanation, as opposed to the mere description masquerading as 'explanation' that is so common in personality psychology.

This complexity is seen in Revelle's (2007, p. 37, our emphasis) statement:

> "Personality is an abstraction used to explain consistency and coherency in an individual's pattern of affects, cognitions, desires and behaviors [ABCD]. What one feels, thinks, wants and does changes from moment to moment and from situation to situation but shows a patterning across situations and over time that may be used to recognize, describe and even to understand a person. The task of the personality researcher is to identify the consistencies and differences within and between individuals (what one feels, thinks, wants and does) and eventually to try to explain them in terms of a set of testable hypotheses (why one feels, thinks, wants and does)."

The simple-sounding 'explain them' has a heavy inferential load to bear.

Of course, science must start with atheoretical observational regularities. Without these, there is nothing to explain. But observation of a regularity, even when it allows some prediction by extrapolation, is not a true explanation – think back to melanin concentration and apparent eye colour. In biology generally, and personality neuroscience in particular, regularities in ecological, social, and behavioural observations can, *in principle*, be explained by lower-level constructs, but that is not enough: also required is emergence and, ultimately, evolutionary argument for full explanations. In personality psychology, a complete evolutionary argument requires linking of molecular, neural, and "subcognitive" levels with higher-level cognitive, all the way up to a gestalt perception of superfical (nonetheless important) regularities – and these regularities need not be universally as opposed to locally consistent. Extraversion, for example, may be *expressed* differently across cultures (and certainly across species), yet still retain the same function(s) and, we would argue, conserved neurobiology.

These considerations compel us to argue that there is a need to understand personality via explanation proper, rather than simply 'explaining away' via tautology. The biological approach to traits is the best lodestar in this quest. But, as when the polynesians set out to discover new islands, the discovery of new psychological explanations is likely to require additional heuristics (equivalent to direction of birds flying with food in their beaks, and particular forms of cloud – New Zealand's Māori name Aotearoa means "Land of the Long White Cloud").

For example, a murmuration of starlings (Fig. 2A) has a complex form with ever-changing dynamics. Why is this? Careful analysis of 3D measurements of the individuals in murmurations by the STARFLAG project (see King and Sumpter, 2012) allowed simulations of the birds' individual behaviour that tested various explanations for the observed form of the flock-level murmurations. The simulations found that each bird reacted quickly to changes in position of a fixed (6−7) number of its neighbours using *topological* rather than *metric* rules. Importantly, addition of an attacking hawk to the simulations suggested that topological rules maintained flock cohesion better than metric ones (Fig. 2B).

The key point is that the simulation provides an *emergent* explanation of flock behaviour based on a simple set of rules that controls each bird *individually*. The murmuration is not a result of any higher-order flock-level control. Nor is it immediately deducable from what is known about individual starling flight control. Note that the results also imply a second kind of explanation of the behaviour – in terms of evolutionary function.[3] "Chance" genetic mutations in individual birds appear to have combined with the "necessity" (Monod, 1972) of avoiding hawks to instantiate the individual-level behavioural rules that then cause the flock-level murmuration to *emerge* spontaneously (Fig. 2C). But, despite individual-level genetic control, the murmuration is still a species-level trait: It depends on the nature of individual birds but can only be expressed by an entire flock in an ecology – and it is the behaviour of the flock that supplies the adaptive advantage that dirves selection.

[3] Strictly, teleonomy. That is, the historical functional path taken by the relevant series of adaptations during evolution of the current phenotype. Current function or "purpose" are a convenient way of thinking about this provided they are not taken to imply teleology.

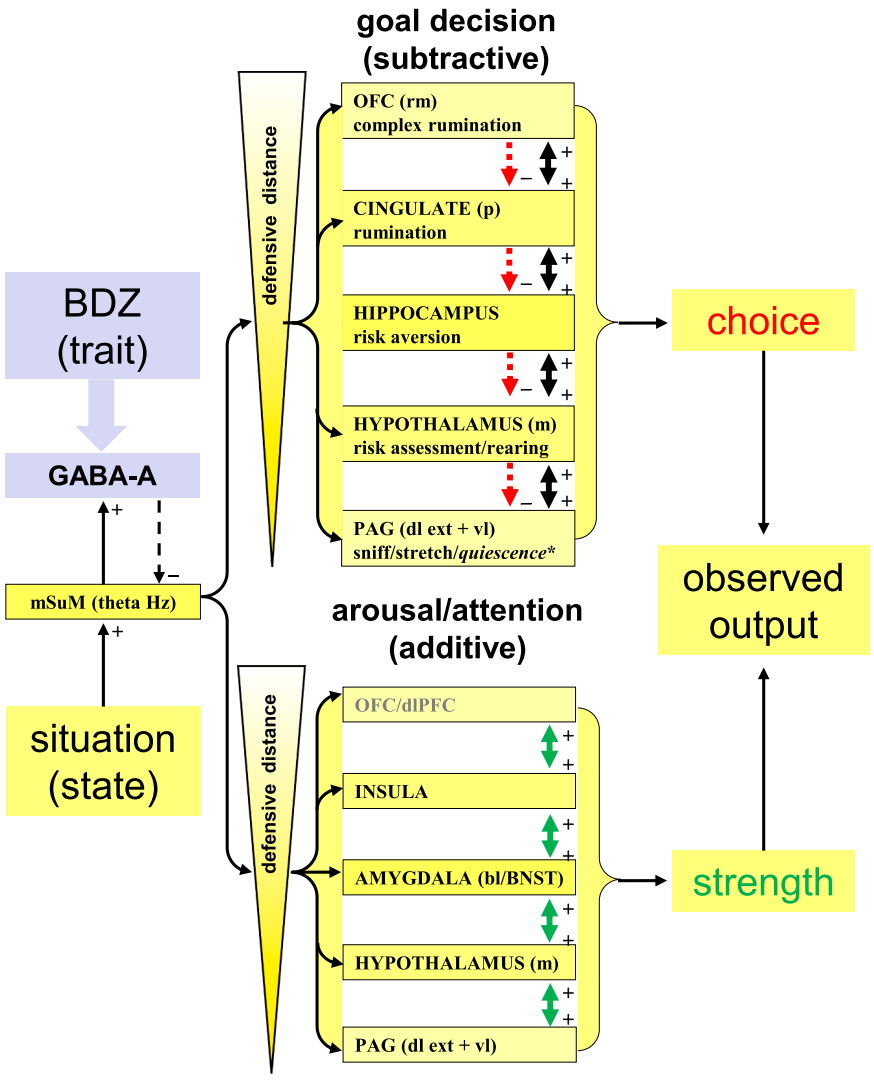

Fig. 6. Modulating input to and generation of output from the GIS. For personality theory the key part is the left-hand side. This gives known neural details (McNaughton and Gray, 2024) of the "theta" input to the entire GIS, which controls its sensitivity to conflict (which is altered by specifically anxiolytic drugs and, in trait terms, endogenous benzodiazepine receptor ligands). The central panel shows the known machinery that control the specific behaviours output by the system at specific defensive distances. Anxiolytics alter defensive distance rather than blocking any specific behaviours. Decision and arousal combine to determine the nature and strength of the output behaviour (Gray and Smith, 1969). Critically, both reward sensitivity and punishment sensitivity are individual trait values and so the balance point between currently experienced reward expectation and punishment expectation (and so when conflict is experienced) varies across individuals for any given set of external reward and punishment probabilities (Pandey and Osinsky, 2025). BDZ = benzodiazepine; b = basal; d = dorsal; ext = external; GABA-A = γ-amino-butyric acid receptor A; l = lateral; m = medial; mSuM = medial supramammillary nucleus; OFC = orbital frontal cortex; p = posterior; PAG = periaqueductal grey; PFC = prefrontal cortex; r = rostral; vI = ventral internal.

Note that the simulations provide a reductive explanation of the flock behaviour – albeit only one level of explanation down (the cognitive level, Fig. 2C). These cognitive processes themselves can be treated as explicanda, predicted by neuromorphic (brain-like "neural network" computer models) as in Marr's (1976); (1982) computational approach to vision (McNaughton, 2020). Neuromorphic processes can, in turn, be explained at the true neural level. Note that neuromorphic models are often derived upward to *test* predictions from, at least large scale, neural observations. They are seldom derived downward via computer modelling of cognitive observations. For a genetic-level explanation of flock-level behaviour all of the intermediate levels must be linked to each other. Each lower level set of observations provides an explanation of the next level up – but with appropriate inclusion of emergent properties. Indeed, it is emergence (in a sense) that defines where a level-break occurs. It is when you need emergence in your explanation that you need a new level of explanation.

Reductive explanation, though, is often misunderstood as being completely reductive and absolutely predictive. However, explanations of social behaviour purely in terms of quantum mechanics are clearly impossible to use. Indeed, neuromorphic models (with cognitive consequences and neurally constrained machinery) are often needed because an explanation of a cognitive process in terms of thousands of individual neurons is too complex. (It is worth noting that neuromorphic models will often allow characterisation of important emergent features.) So, we need a scaffolding approach; where elements of each level explain (coupled with emergence at that level) the level above and, in turn, are explained by (in the sense of being consistent with) the level below. Neither level will be pre-eminent. It is their combination as explanation and explicandum that provides substantive *explication*.

This stacking of levels is pre-eminent in personality and social behaviour. What we view as complex social behaviours can be reduced to individual actions. In this way, personality psychology is crucial to understanding all forms of collective behaviour. Even the forms of behaviour studied by sociology may be reduced to these lower-level

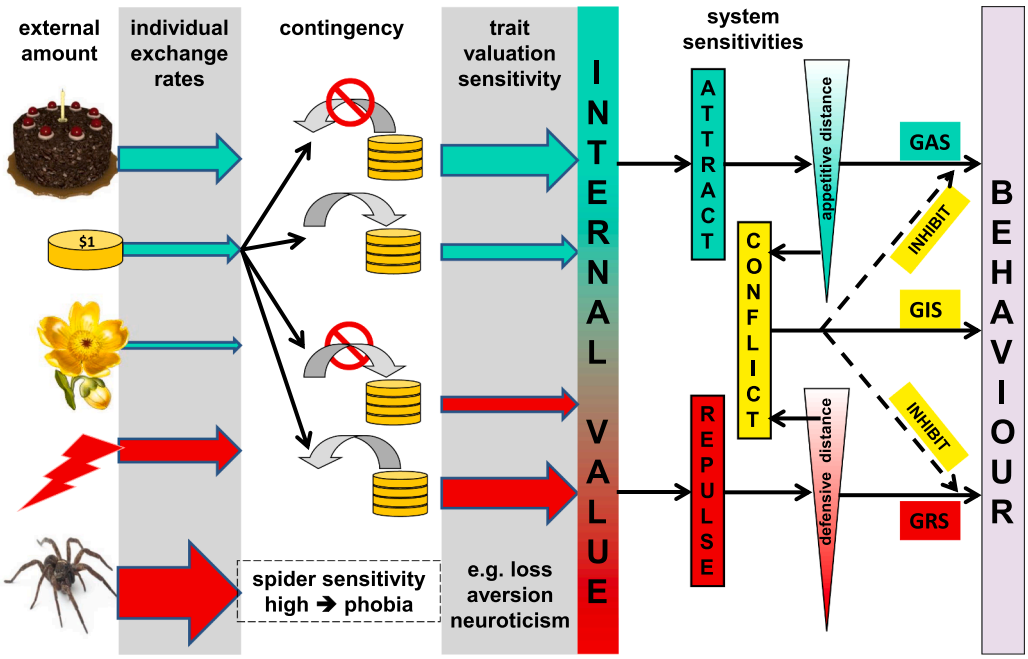

Fig. 7. Trait sensitivities as modifiers of the relations between external amount, contingency, and value. An external item will have a specific amount (e.g., one entire cake) that, together with the current level of drive (which acts like a currency exchange rate) for that kind of item for that person, determines its primary internal value (thickness of arrows in first column). As shown for the case of $1, this interacts with whether the item will be gained or lost to determine the direction and size of its internal value as ultimately measured by the effect on behaviour. The direction of this effect is reversed if the gain or loss is omitted. Loss (removal from a store of items) is most easily controlled with money but will also occur when, for example, one rat steals the food from another rat. Trait sensitivities can be input-specific (spider); valuation specific (loss aversion); or reinforcer-system-specific (GAS = attraction, GRS = repulsion, GIS = inhibition/goal conflict). Input to the GIS from GAS and GRS produces output when defensive and appetitive distances are similar and blocks their behavioural output, substituting it with appropriate conflict resolution behaviour. The generation of behaviour by the systems is shown in Fig. 8. Figure based on McNaughton et al. (2016) with permission conveyed through Copyright Clearance Centre, Inc.

causal personality factors and emergent processes. But there is also vital regulatory input from environment factors, such as government structure, social attitudes, and legal constraints. These can be seen as cybernetic adjustments controlling complex social behaviour.

We are not arguing for personality-based explanations of complex social behaviour. As with quantum mechanics and cognition, more appropriate levels of explanation are available. For example, "When people's personal identities are salient, they act according to their own idiosyncratic norms, values, and morals (more broadly, their personality). [But] when their social identities are salient, they act in accordance with the norms, values, and morals that define their identified category, and this behaviour is just as normative as that when their personal identities are salient" (Ng and Platow, 2024, p. 307). So, we think personality theory can inform social explanations especially when personal identity is salient and, importantly, constrain interpretations that are not consistent with what is known of underlying personality.

6. What do we mean by "a trait" in psychology?

We have already considered the definition of a trait at the biological level; now we turn to what we mean by a "trait" at the psychological level – where there are clearly different emergent properties to consider.

Let us start with the classic answer provided by Allport (1931) to the question, "What is a trait of personality?". He proposed eight criteria:

1. They have more than nominal existence – that is, they exist in their own right.
2. They are more than generalized habits.
3. They are dynamic – and not static.
4. Their existence may be established empirically, or at least statistically.
5. They are only relatively independent of one another.
6. Psychologically considered, they are not the same as a moral quality.
7. Acts, and even habits, that are inconsistent with a trait, are not proof of the nonexistence of the trait.
8. They may be viewed either in the light of the personality which contains them, or in the light of their distribution in the population at large.

We would add that a psychological trait can most generally be seen as *being* (as with height) or *producing* (as with trait blood pressure) consistent behaviour or consistent patterns of behavioural reaction. We emphasised for blood pressure (Fig. 1) that, even in this rather simple medical case, "trait blood pressure" is a construct – essentially a constant multiplier that mediates, for a specific individual, the relationship between rapidly changing energy load and resultant state blood pressures. In this sense, blood pressure per se is not the 'trait', but the trait reflects other aspects of the physiological system (particularly aldosterone) that controls (and *produces*) the state blood pressure giving it a relatively stable sensitivity, or *being* in the sense defined above.

We can use a similar model for psychological measures. Whereas increased load can cause increased blood pressure, the appearance of a predator (increasing the risk of damage) can cause increased state "fear". That is, it systematically generates behaviours such as fight, flight, freezing, directed escape or avoidance; with the specific behaviour observed depending on the cognitive construct of defensive distance (Blanchard and Blanchard, 1989, 1990). It is the set of behaviours as a whole, with the shared function of defence, that is identified with fear. No single behaviour is diagnostic. Likewise, variations in individual trait fearfulness (i.e. sensitivity to fear-generating stimuli) do not cause specific behaviours, rather they alter the defensive distance experienced for any specific physical distance. Note the fearfulness per se does not, by itself, cause fear. Instead, it acts as a multiplier of the effects of any specific fear-inducing stimulus. For state fear to occur there must be an

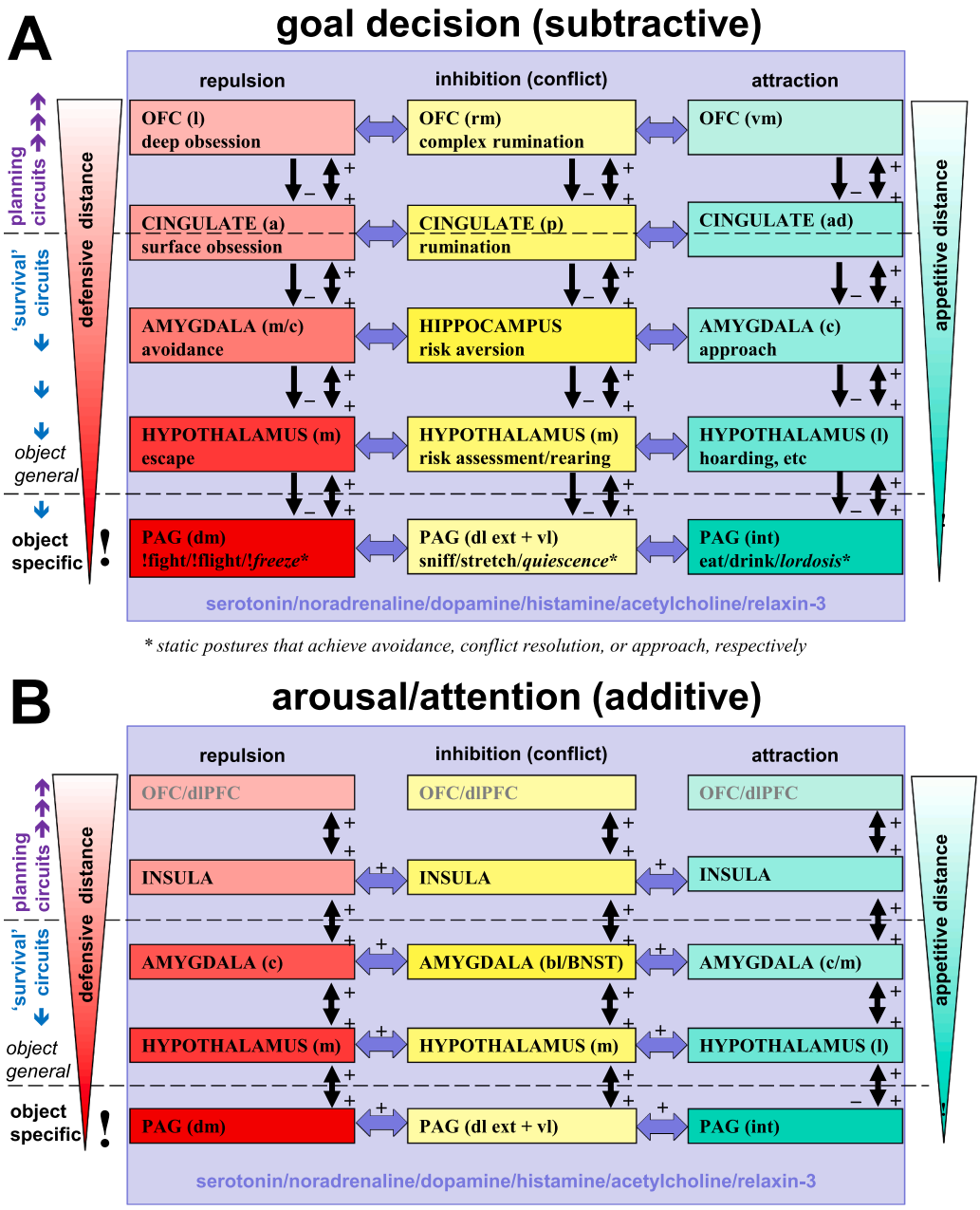

Fig. 8. Hierarchical organisation of goal control by the GRS (repulsion), GIS (conflict), and GAS (attraction) systems. For details of survival circuits, see Chapter 6; for planning circuits, see Chapter 11. Control of subgoal sequences, and conversion of goals to rules, actions, and acts is carried out by non-orbital frontal cortex (e.g., Fine and Hayden, 2022). A. Decision circuits – where effects of attraction and repulsion subtract from each other; and where conflict can block the output from both and replace goal direction with risk assessment. B. Arousal/attention circuits – where adding attraction to repulsion, or vice-versa can result in a more energetic goal-directed output; and where conflict (essentially equal attraction and repulsion) increases arousal further. Figure and legend from McNaughton and Gray (2024) with permission via author's contract. Abbreviations as for Fig. 6 plus: c = central; int = internal.

internal cognitive construct (real or imagined) to be fearful of. That is, the simple presence of snake will not *per se* induce fear (especially in a well-trained snake handler). But if a piece of hose in the bush is perceived as a snake about to strike, that will generate fear. Likewise, with obsessive compulsive disorder fear will be generated by the perceived risk of infection (for which there is necessarily no detectable external stimulus) even when the hands have been sufficiently washed that a person without obsessive compulsive disorder would be sure there is no danger.

However, it should be noted that the rodent approach described is based on a large mass of systematic ethoexperimental analysis of state behaviour under naturalistic laboratory conditions. That is, the description of consistent behavioural patterns preceded any attempt to look at individual differences. Human personality psychology has attempted to find less laborious, language-based methods of arriving at such systematic descriptions – using questionnaires for fundamental data collection and standardisation. This approach has been responsible for the focus on systematic (observed) behaviours as "the trait" of personality – it serves, though, to conceal the underlying causal trait(s).

In personality psychology, a potential advantage of the use of words to capture current and/or past behaviour is that phrasing can, in principle, capture the nature of the responses to trait features of behaviour filtering out state noise. "Do you often …" should bypass reports of current state – and some clinical questionnaires have tightly matching state and trait forms of question that can give very different answers at the same point in time (e.g., Spielberger et al., 1983). This aspect of

questionnaires bypasses a major problem in animal work. In addition to the fact that measurement of ethologically-defined traits is not easy, the bulk of rodent measures generally confound state and trait components (as with blood pressure, Fig. 1B).

Questionnaire methods normally used can be split into two basic types: lexical/descriptive and what we will call lexical/rational; with a third type: an anchored/rational variant linked to how RST-based questionnaires are constructed. We consider these approaches to explore the distinctions between descriptive and causal traits.

6.1. The lexical/descriptive approach

The lexical/descriptive approach starts with words and sees these as capturing traits because of their evolution within the language – reflecting evolved behaviours (e.g., sociability). (An example of lexical evolution, here, could be the creation of the word "murmuration" that labels the flock-level trait shown in Fig. 2A.) The words (considered as items) are not chosen to map to prior supposed theoretical constructs; explanations of the clustering of the words must be derived *post hoc*. We can view the lexical/descriptive approach as using a large pool of observers (and a very large period of time) to extract and then label the consistent patterns of ABCD, as suggested by Revelle (2007). This descriptive approach has proved highly compelling in summarising individual differences in human behaviour. It has a strong presence in the descriptive trait approach.

We can see the power of the lexical/descriptive approach in its assumption that "those individual differences that are of most significance in the daily transactions of persons with each other will eventually become encoded into their language. The more important is such a difference, the more people will notice it and wish to talk of it, with the result that eventually they will invent a word for it" (Ashton and Lee, 2005 citing; Goldberg, 1981, pp. 141–142). This makes the lexical/descriptive approach particularly attractive for traits that *emerge* at the social level (i.e., during social interaction; e.g., trait social anxiety defined as an individual's sensitivity to state social anxiety – with excessive levels delivering social anxiety disorder). Note that our murmuration example is at the next level up from this. It is a species, not an individual, trait. Here we are talking about individual traits conditional on their being embedded in a social situation.

There is a significant scientific problem here. To see this, let us return to eye colour. We noted that human categorisation (which is based on the physiology of the human visual system, not on the reality being processed) produces verbal labels that are qualitative and nominally discontinuous (blue, green, brown, and several more). But the endogenous trait giving rise to the colour, melanin concentration, is a simple unidimensional value. With lexical measures, as with colour names, the subjective human viewpoint is the starting point. So clearly, we will need care in mapping the verbally-derived construct to underlying (causal) reality, and the resulting trait. Equally important is the fact that many extracted traits have no objective referent (like the actual colour of the eyes). Thus, the nature of the trait must be inferred post hoc once a lexical cluster has been discovered.

That said, it is hard to see how one can arrive at an explanatory framework for a high-level emergent construct of the lexical type without an at least initial lexical pattern analysis. Descriptive statements seem needed first to derive descriptive traits. To reject the lexical/descriptive approach is to reject the fundamental explicanda that only it can easily extract – at its own level of representation, it is valid. But, at other levels, this may not be the case; and herein lies the problem of descriptive versus causal traits. It involves observations of individuals by the population, in the way that the lexical/rational approach (next section) often involves observations of individuals by themselves or by others, and ethological approaches involve observation of individuals by experimenters (ideally with inter-rate reliability checked). In all these cases, the behavioural observations must be processed further to derive assessments of traits. In explanatory terms, these derived descriptive traits must be at a deeper level than the observations supplying the recorded regular patterns, otherwise we achieve only tautology.

The observed factor analytic conjunction of multiple words (e.g., organized, thorough, disciplined, responsible, decisive, cautious, studious, practical) into a scale implies that they share a common element at a deeper level "which we happen to call Conscientiousness" (Ashton and Lee, 2005, p. 15; who see also for general rebuttal of objections to the lexical approach). This deeper level is strengthened where there is similarity of such factors across a range of languages. But note their "we happen to call …". The labelling of such scales can involve heroic assumptions about what in reality has given rise to the observed inter-correlations. This is where deeper explanation becomes important if we are to truly understand causes, emergence, and expression of those elements that define personality.

The lexical/descriptive approach can be argued to provide some explanatory power rather than reflecting simply observational regularities. "When we learn that a set of characteristics load together on a factor, we do learn something about the psychological meaning of those characteristics—something that was not necessarily predicted prior to the discovery of the factor" (Ashton and Lee, 2005, p. 14). However, in our view, this conclusion contains a dangerous scientific aspect: such pattern generalisation depends massively on semantics, and there is no strong reason to assume they map onto underlying causal factors. It may be predictive of other patterns in a language but is not necessarily predictive in terms of explanation from lower levels. Nor does it allow identification of the causal trait(s) in the person that gives rise to the observed population pattern. This must involve not only lower-level constructs explaining the higher level but identification and explanation (as with the murmuration) of any emergence. In this context, emergence is not a get-out-of-jail-free card. It has lower-level causes and with the starlings, at this lower level, these are clearly (Ballerini et al., 2008) of one type (topographic) not another (metric).

Ashton and Lee (2005) start with essentially arbitrary semantic entities and use factor analysis to extract lexical coherence. When a set of characteristics load together on a factor this is taken to reflect personality as contained in questionnaires. This approach gives priority to the factor analytical relationship of semantic entities – typically, this is assumed to be personality as contained in questionnaires. However, even in the lexical domain, the initial factor solution may not contain what are ultimately taken to be the purest distinct factors. There are an enormous number of solutions based on axis rotation. Varimax solutions (that place axes through clusters of items) are often preferred as varimax is not entirely arbitrary. But even then, one must ask if an orthogonal or oblique solution is more appropriate for one's current purpose.

These issues do not go away if we start with a more anchored pool of standard psychiatric ratings of a relatively homogenous group of neurotic individuals – rather than experimenter-constructed items and a general population. Although initially based on psychiatric signs and symptoms, this variant of the lexical approach was adopted by the extension of Hans Eysenck (Eysenck, 1944, 1947) work in the production of the Maudsley Medical Inventory (MMI) as a neuroticism questionnaire. He provides detailed arguments for care with rotation and, for his data, used a separate "criterion" variable (and inclusion of a range of neuroticism-related tests) to align his subsequent rotation (Eysenck, 1950). He then developed the Maudsley Personality Inventory (MPI).

> "The E (extraversion) and N (neuroticism) scales of the MPI were derived from rather elaborate procedures involving item analysis and factor analysis of other personality inventories, principally the Guilford inventory of factors STDCR and the Maudsley Medical Questionnaire. The two scales, E and N, have high 'construct validity,' that is, the items making up the scales are highly correlated with the factor they are said to measure and they have insignificant correlations with other factors. The items have been selected so as to minimize the correlation the E and N scales. The two factors are thus represented as orthogonal, i.e., uncorrelated with one another"
>
> Jensen (1958, p. 315).

Subsequently, The Eysenck Personality Inventory (EPI) departed from the MPI by: rewording items; removal of items causing a small correlation between N and E (supposedly removed previously in the MPI); and adding a Lie Scale (Eysenck and Eysenck, 1964, p. 5). The Eysenck Personality Questionnaire (EPQ) supplied a major revision (including adding a Psychoticism scale) but despite claims of factor equivalence has an E scale that is almost purely sociability, while the MPI E scale is a mix of impulsivity and sociability (Rocklin and Revelle, 1981). The revised EPQ-R was a psychometrically improved version of the EPQ that appeared in 1984 (Eysenck et al., 1984). Direct comparison of Eysenck's different versions (judging them against the factors extracted when the MMI,[4] MPI, EPI, and EPQ-R are all included together in a single varimax analysis) found high (>0.85) positive correlations for all E and N scales with the relevant varimax factor and small (>-0.32) negative correlations of E with varimax N, and of N with varimax E. "Overall, the EPQ-R seems to be the most valid single measure of the PEN model" (Ruch et al., 2021, p. 1; correlations with relevant varimax factor, E = 0.93, N = 0.91).

This history shows that logical development can be needed to remove statistical and linguistic problems with scales that are primarily lexically-derived. In that sense the EPQ-R scales appear thorough and reliable. Eysenck's scales were in a sense anchored in clinical reality and so less problematic at the start than others that start without such an anchor. But, despite a strong interest in the underlying biology, Eysenck only attempted to explain the causal basis of these descriptive traits after finalizing the initial scales lexically and statistically (e.g., Eysenck, 1957, 1967) and never determined a definite substrate for the factors, nor then develop scales anchored to the biological reality. This raises questions both of factor content and of rotation to appropriate scales.

In terms of factor content, we can compare Eysenck's E and N scales with those of the NEO-PI (e.g., Draycott and Kline, 1995). This shows substantial overlap of the scale spaces. But when we compare the location of the scales within their common space (Supplementary Figure 1) there are significant discrepancies with three dimensions required to display the two nominally common factors. In terms of rotation, "as pointed out by Eysenck (1965), the trait of susceptibility to anxiety, measured by the Manifest Anxiety Scale (Spence and Spence, 1966; Taylor, 1953, 1956) is loaded on both the Eysenckian dimensions of introversion and neuroticism, though the correlation with neuroticism is somewhat higher" (Gray, 1970a, p. 252).

So, in the "neurotic" population that Eysenck started with, Trait Anxiety (Gray, 1970a; Spence and Spence, 1966; Taylor, 1953, 1956) occurs as a cluster that is located within the neurotic introvert quadrant of the two dimensions. So Gray, at this point, proposed a rotation of the scale axes (essentially using a cluster of anxiety disorder cases within a group of 700 male soldiers referred to the Mill Hill Emergency Hospital, as a varimax target). After this rotation, he renamed the resultant rotated dimensions as punishment sensitivity and reward sensitivity. This dramatically changed the target for the search for the underlying biology of the factors.

Unfortunately, while doing this, Gray provided arguments for an identity of a "punishment" factor reflecting sensitivity of his BIS with this new "punishment" factor derived from neurotic introversion. This conflated at least three meanings of punishment, in one case within a single sentence: "a punishment [1] mechanism for passive avoidance [and] a separate punishment [2] mechanism for organizing the unconditioned response to a punishment [3]. We shall call this the "fight/flight" system" (Gray, 1971, p. 194). Note that punishment [1], punishment [2], and punishment [3] are all required to allow his inclusion of all the then "anxiety disorders", with, e.g., punishment [3] required to include OCD. However, this psychiatrically recognised group included many (e.g., OCD) that are not sensitive to anxiolytic drugs and so not dependent on an oversensitive BIS. Thus the resultant rotated Eysenck-space axis did not in fact map purely to Gray' s anxiolytic-defined BIS but included his FFFS (for detailed exposition on this, see McNaughton and Corr, 2019).

Let us consider one consequence of this approach of prime importance. It is known that there is a strong relationship between personality and psychopathology, especially involving neuroticism. It is highly limited in explaining the scientific nature of neuroticism-related psychopathologies purely in descriptive terms, even with the aid of complex statistical models. To understand the causality of psychopathologies, and thus recommend viable treatments, we need to know the neural mechanisms, and how they emerge to the higher-level symptoms (emotions, behaviours and motivations).

In conclusion, the lexical/descriptive approach may be a good place to start but, we argue, not to end. In one sense it may not even be a good place to start, as it is hard to decompose a high-level construct into underlying mechanisms, but in many cases may simply be the only place available. It has the potential to misdirect the scientific quest. However, if we only start at the mechanistic bottom, we may be unable to see the wood for the trees and we may fail to compose the needed higher-level constructs. In relation to levels of explanation, then, it may be important to try and cycle between the two (sandwiching, where necessary, a neuromorphic level).

6.2. *The lexical/rational approach*

The lexical/rational approach treats the person answering a questionnaire as a simple observer. The questionnaire and its related traits are, of necessity, descriptive in nature. In principle, this is not much different from the ethologist recording the behaviour of baboons; though it lacks the objectivity of ethoexperimental analysis in a laboratory (Blanchard and Blanchard, 1988). The rational approach generates initial pools of items in terms of the experimenter's perception of their capture, in ordinary language, of a concept of interest. Validation of the resultant questionnaire then tries to ensure face, content, criterion, and construct validity; reliability can be assessed by intra-rater (test-retest), inter-rater (Ranganathan et al., 2024), and some form of internal consistency (e.g., Cronbach's alpha).

Of most concern, here, is that verbal items may not, in fact, access the supposed real-world construct that the questionnaire was designed to measure (i.e., the causal traits) – or that the labelling of the construct does not reflect the criteria used for questionnaire construction. These are the well-known jingle-jangle fallacies that reflect these problems. This is best seen with two examples – these are important in showing the clear-cut differences between causal and descriptive traits.

One physiology-related example is the labelling problem of the Autonomic Perception Questionnaire (Mandler et al., 1958). This instrument was, in part, designed to assess those with anxiety reactions but it explicitly focussed on participants' *perception* of their level of autonomic activity. (Given that anxiety is perceptible, this was not an unreasonable assumption.) Original validity testing involved selection of high and low scoring groups in terms of their statements about their perception of such autonomic changes as they had; and *who were further selected on their interview reports of a stressful situation.* "High perceivers showed significantly greater autonomic reactivity than low perceivers; … [with] positive correlations between scores on [the APQ] and other paper-and-pencil tests of anxiety reactions" (Mandler et al., 1958, p. 372). That is, it appears that those reporting good autonomic perception were in fact simply those experiencing high levels of autonomic activity. So, their report is supposedly of accuracy of perception of any autonomic changes but there was no direct test of this accuracy of perception.

Problems arose with the perception accuracy interpretation. "Several studies have employed the … APQ, as a measure of subjects' awareness of their Galvanic Skin Responses or heart rates. [But,] the expected

[4] The MMI is called the Maudsley Medical Questionnaire by Ruch, W., Heintz, S., Gander, F., Hofmann, J., Platt, T., Proyer, R.T., 2021. The long and winding road: A comprehensive analysis of 50 years of Eysenck instruments for the assessment of personality. Personality and Individual Differences 169, 110070.

positive relationships between [APQ] and autonomic control was not obtained" (McFarland, 1975, p. 402). Direct comparison of APQ scores with physical tests of actual heart rate perception found no relationship (McFarland, 1975; Whitehead et al., 1976). "The APQ is not a valid measure of ability to perceive heart beat" (Whitehead et al., 1976, p. 177). Here the problem appears to be with the ambiguity of "autonomic perception". What seems to have been tested for and validated is whether someone reports or does not that they experience autonomic changes – with reporters generally having higher actual reactivity not greater sensitivity to change independent of level. But what then appears to have been taken by some as the meaning of the APQ value is the sensitivity to change, not the size of the amount usually experienced. Amount of reaction would, of course, be expected to relate to the processing of anxiety (with unknown causal direction) giving the APQ an undeserved appearance of validity.

A similar conclusion can be draw about the original scale nominally constructed by Carver and White (1994; C&W) to assess sensitivity of the BIS. As noted by McNaughton and Gray (2024, p. 399), "The items of the C&W 'BIS' scale ***do not*** *relate to behavioural inhibition*, nor to punishment as a source of passive avoidance, nor to goal conflict. Instead, they focus on worry (40%), anticipatory fear (30%), and stress (15%)". That is, the higher-level emergent properties of the BIS. We start to see the logic and associated problems when we ask how C&W (1994, p. 322) constructed their "BIS" scale.

> "In approaching the assessment of BIS sensitivity, we attempted to create statements that reflected a concern over the possibility of a bad occurrence ('I worry about making mistakes') or a sensitivity to such events when they do occur ('Criticism or scolding hurts me quite a bit'). All potential BIS items thus referenced potentially punishing events and asked how people respond to them. … Rather than … assessing BIS sensitivity in terms of reports of behavioral reactions, we focused instead on the experience of anxiety per se in situations in which there are punishment cues."

This approach is in contrast to the Gray-Wilson Personality Scales (GWPS; Wilson et al., 1990) that are specifically behaviourally based, but its factor structure does not reflect the underlying RST constructs (Corr, 2016). The C&W BIS/BAS scales have proved enormously popular. As of February 2026, they have been cited over 11,000 times; and, like the APQ, they clearly measure something of interest. However, the C&W BIS has not been validated biologically in terms of its original neuropsychological base. A review of fMRI studies only found RST scales (like C&W) based on Gray (1982) rather than the revised Gray and McNaughton (2000) and concluded that

> "psychometric and experimental limitations still hamper the differentiation of the BIS and FFFS systems in their activation of deeper brain networks. Future studies need to include revised RST scales that separate the BIS and FFFS and implement more rigorous tasks that allow for the examination of each system both independently and codependently" (Standen et al., 2022, p. 395).

Crucially, like neuroticism, C&W BIS appears equally related (Johnson et al., 2003) to all the "anxiety disorders" (generalized, panic, social, and post-traumatic) and to the "depressive disorders" (major depression, dysthymia). The same appears true for, e.g., the Sensitivity to Punishment scale (see Barrós-Loscertales et al., 2006, p. 1011) despite its relation to hippocampal volume and potentially because of its overlap with Harm Avoidance (see Barrós-Loscertales et al., 2006, p. 1012). Unfortunately for Carver and White, the neural BIS is originally defined in terms of sensitivity to specific anxiolytic drugs – which rules out the bulk of C&W disorder correlates, which are only sensitive to panicolytic antidepressants. So, again like the APQ, it seems mislabelled (causing substantial confusion) and is theoretical misleading, especially as it has been so widely used in BIS studies – it is perhaps better viewed as a worry scale.

So, to be a reliable, face-, superficial content-, internal-valid measure does not mean a test has criterion validity. Though, of course, the real failure in the above two cases is linked to misunderstanding of the criteria for the content. But, how do we ensure criterion validity? One can of course, as we just have, check this *post hoc*. With both the APQ and C&W BIS, the problem is that true validity was never checked at all during test construction – with physical measures being required in both cases and the test failing when these are applied.

One potential solution to this problem is to use rational scales constructed employing criterion anchors (Eysenck, 1950). As an example, Hans Eysenck (1944) started with a particular chosen distinct population – soldiers referred to Mill Hill Emergency Hospital with "neurotic disorder". The soldiers were those referred to the hospital, not recruited by him, and he made no attempt to differentiate diagnoses. Ratings and tests were a subset of 39 items from ~200 items recorded by psychiatrists (i.e., not self-report) and, importantly, were provided by more than 12 psychiatrists, ruling out individual rater bias in determining the factor structure. Note that the form of the items was determined by the hospital's normal practice. Selection of items from the pool was on the basis of maximum objectivity and for having "some definite psychological bearing on the illness and the personality of the patient" (Eysenck, 1944, p. 852).

Eysenck used factor analysis to ask what underlying factors the patients had in common. Some items had no obvious statistical relationship to others (age, bombing and exposure, alcohol, headache, degraded work history, etc.). However, overall, four factors could be recovered (with 14%, 12%, 8%, and 6% of the variance respectively for a total of 40% over all the items included). In Eysenck's (1944, p. 854-855) view, "quite clearly, the [first] factor is one of 'neuroticism,' or 'lack of personality integration'. … The second, bipolar factor … we consider … to be identical with the introvert-extravert dichotomy". The other two factors appeared to represent hypochondriasis and a distinction "between the stupid, drunken, shiftless social misfit on the one hand and the 'psychological conflict' group on the other" (p. 857) – and were treated as irrelevant for his purposes.

What is important to note here is that neither of the first two factor scales, as such, reflected neurosis. Indeed, Eysenck saw them as general personality factors that also emerge in "studies of young and old, normal and abnormal, human and animal" (p. 857). While not representing disorder, neuroticism (especially when coupled with introversion) appears to be a general risk factor for the development and perpetuation of a range of internalizing disorders (Andrews et al., 1990; McFarlane, 1989), especially when combined with high levels of distinct disorder-specific traits (McNaughton and Glue, 2020; Fig. 5). Similarly, agentic (as opposed to communal) extraversion "tended to have positive associations with psychopathology; it displayed particularly substantial links to indicators of mania, narcissism/narcissistic personality disorder, and traits related to externalizing" (Watson et al., 2019, p. 777). However, as we noted in the previous section, while Eysenck's item pool was arrived at somewhat rationally, he did not have a clear biological anchor for his neuroticism trait and made no attempt to use one in the subsequent, essentially lexical, scale development.

Note that, as with our two previous cases, we have reason to see the rationally-derived scales as measuring traits that exist in more than a nominal form (Allport, 1931). But what is needed to identify the traits is reductive explanation. It was not until the development of RST, starting with Gray (1970a), that underlying causal traits were taken seriously, and a clear separation was made between them and descriptive traits, hence the need to 'rotate' Eysenck's Extraversion and Neuroticism measures (for an exposition, see Corr, 2008).

6.3. The anchored/rational RST approach

So far, we have looked at different primarily top-down approaches to psychological trait measurement. The lexical approach is almost purely top-down. It has high-order presumed-trait constructs inferred from the evolution of the words of a language. These are not even cognitive-level

constructs, they are a product of what can be thought of as a group mind. The rational approach *does* start with *nominally* lower-level theoretical trait constructs (e.g. autonomic perception). But even these are not cognitive-level in explanatory terms. They are usually derived intuitively, not from biological measurement; and the rational approach then uses them in a superficial, linguistic, interpretative way.

With RST, however, the starting point is neurally-defined *state* systems (Gray, 1969; Gray and Smith, 1969). The bottom-up idea, that each state system has a trait sensitivity mediating the effects of its inputs, was then added (Gray, 1970a). Biologically, this view of traits is a tautology except in respect to the large number of the neural structures that in Gray's proposal are dependent on a single trait and are not pre-defined in trait terms. Given that the proposed state systems are biologically defined, they have the uncomfortable property that their presumed psychological nature, in human verbal terms, can change based on new data (hence the 1982, 2000, and 2024 editions of the state theory, and the need for the later parts of this paper). But that is how science operates – the theory must fit the facts.

The neural basis of state RST has evolved (or at least expanded) – but what is often missed is that this only affects a few detailed behavioural predictions. The *core* functional elements at the trait level can be treated as invariant. But the psychological naming shifted from, "behavioural inhibition" (Gray, 1976) to "goal inhibition" (McNaughton, 2020; McNaughton and Gray, 2024) as the biological function of those *same* neural system became clearer. Note that the change in psychological name does not change the neural identity nor the presumed functions of the original system. In the same way, changing the name of some species from geranium to pelargonium (in 1789) improved botanical classification but did not change the identity of the plants themselves. (This particular renaming still, in 2026, meets resistance.)

What has been lacking is good measures of trait sensitivity. The state RST theory can suggest a plethora of behavioural measures that should capture trait sensitivity. However, in practice, such measures confound state and trait variance and, in any case, should be administered as a battery to obtain, e.g., GIS sensitivity as opposed to some related higher order (e.g. neuroticism) or lower order (e.g. obsessionality) trait component. In that sense, behaviours are just like questionnaire items – but for non-linguistic folk.

A solution to this problem could be the derivation of truly rational RST questionnaires. But these obviously need to avoid the use of inappropriate items (with the C&W BIS that we have already discussed as a major warning). We argue that such questionnaires will need to include appropriate neuropsychological anchoring (both during and after construction). So, these would be "anchored/rational" not "lexical/rational" in origin. Note that, with neural systems, the anchoring is stronger than Eysenck's anchoring to a heterogeneous clinical group. His mixed "anxiety disorder" group might have some form of common specific neural dysfunction but this was not demonstrated and would be hard to replicate. This lack of clear neural anchoring drove Eysenck's subsequent unsuccessful search for a neural basis for his factors. Eysenck, here provides an important model for part of the necessary RST processes. Anchoring will be essential, but the clear and steady progress made in the development of what was ultimately the EPQ-R shows that an iterative process will be needed: ensuring comprehensible items in scales of high statistical quality, biologically validated at each stage of development.

In recent years, several "revised RST" scales have been constructed using the lexical/rational method (for some comparative comment see Krupić et al., 2016; Leue et al., 2022). These include: the 'Jackson 5' (Jackson, 2009); Reinforcement Sensitivity Questionnaire – RSQ (Smederevac et al., 2014; see Smederevac et al., 2022 for genetic analysis); the revised Reinforcement Sensitivity Theory Questionnaire – rRST-Q (Reuter et al., 2015); the Reinforcement Sensitivity Theory-Personality Questionnaire – RST-PQ (Corr and Cooper, 2016); and the Hierarchical Reinforcement Sensitivity Theory of Personality Questionnaire – HRST-PQ (Moncel et al., 2023). This plethora of scales and the fact that each measures a somewhat different set of factors (in somewhat different numbers of dimensions) raise questions about their detailed construction and general lack of neuropsychological anchoring.

We will not go into detail here but, as an example of linguistic sources of the problems, let us briefly consider a single item, "fight". This label within neural RST caused many problems during RST test construction. Many experiments simply omitted it, or assigned it to its own distinct scale. The problem arises in the historical labelling of the Fight, Flight, Freeze system. Despite the name, biologically it includes systems controlling escape (with escape being a level above the panic reaction of "flight") and simple and complex avoidance – with neural control way above that of fight, flight, and freeze (McNaughton, 2020). The fight/-flight component refers to behaviour shown only when ultimately close to or contacting a predator. It involves defensive fight in the form of direct screaming attack as a means of escape – it is an acute inter-species action (but may perhaps occur in acute intra-species reactive aggression). It can be contrasted with intra-species fight in the form of social aggression and instrumental aggression; and with inter-species predatory aggression (e.g., the killing that you do if you go hunting or fishing). The different scale systems have then made, in a minor form, the same mistake as is made with the APQ. They have taken the English word and used their interpretation of it for what is essentially lexical scale construction and have not tried to map the scale item to the theoretical *neural* system that it is supposed to assess.

This suggests, strongly, that when comparing the questionnaires we should carefully compare their *items*. We should prefer those that are more clearly objective reports of events and ignore the items (typically a majority) that have complex or bizarre linguistic forms (e.g., "Conscientiously feed parking meter", "Prone to mild addictions", "Careful to complete homework at school", Wilson et al., 1990). We should also prefer those items that clearly converge on the original proposed theoretical elements and could, in principle, be asked of another species. Where a GIS item, for example, references worry, or anxiety, or anything unrelated to goal inhibition, it should be deleted. This approach could allow a rational coalescence of the current RST systems – and be much more functionally informative than any amount of inter-correlation.

As an example of one RST scale system that may be beginning to solve some of the pure-rational problems, let us consider the rRST-Q (Reuter et al., 2015). First, they excluded 'fight' during their scale construction because "in the revised RST *fight behavior is only observable if the distance between predator and prey is close to zero*, leaving no option for flight or freezing. The probability of such situations occurring for human beings is extremely low" (p. 5, emphasis added). This takes the neural theory's definition of "fight" at face value. Further, in the context of the hierarchy of the neural system, omitting one of 3 situation-specific extreme bottom level behaviour items should not disturb assessment of the relevant defensive distance sensitivity and allows for species differences, particularly in testing environment. Second, unlike all the other revised RST scale systems, rRST-Q provides one questionnaire scale for each of the neural systems, again matching theory. Third, they challenged the resultant scales with assessment of a functional genetic polymorphism (rs11174811) on the AVPR1a gene thought to link to anxiety but not fear. This was associated with their rBIS scale scores but not scores on other rRST-Q scales.

Of particular interest, the original rRST-Q scale structure has been replicated (Espinoza Oyarce et al., 2022); and a similar BIS/BAS/FFFS structure (Espinoza Oyarce et al., 2021) neurally validated via *structural* MRI (Espinoza Oyarce et al., 2024).

> "The BIS was negatively associated with hippocampus laterality [i.e. a relatively larger right hippocampus]. At standard significance levels, the fear component of the FFFS was positively associated with anterior cingulate cortex; the BAS was positively associated with bilateral caudate; and the BIS was positively associated with posterior cingulate cortex volume. Furthermore, these neurobiological systems showed distinct patterns of association with cortical

thickness though future work is needed. Our results showed that the neurobiological systems of the revised RST characterized in rodents can also be identified in the human brain."

(Espinoza Oyarce et al., 2024, p. 1)

As noted earlier, one would expect MRI structure to drive traits, while current momentary functioning drives states. That is, a long-term trait can be caused by (or result in) structural change – as with the reduction in hippocampal size with PTSD or major depression (see, e.g., Sapolsky, 2004, p. 221). With fMRI, any background trait effects will be confounded with current state variations. A similar problem arises with EEG as a measure of brain activity. So, MRI may be an at least partial key to solving the neural-questionnaire translation problem.

The rRST-Q has not yet been challenged by the item-by-item comparison we suggest would be useful to coalesce all the scales. Such amalgamation is essential as, if they truly map to the neural theory, they must all measure the same constructs (even if not with the same items). The current chaos is clearly far from this desirable outcome. While rRST-Q is looking hopeful in terms of biological separation of BIS/GIS from FFFS/GRS – and the whole system in terms of at least preliminary mapping to neural volumes – this is clearly a beginning not an end.

The problems with chronic stress (resulting from release of CORT[5] by anxiety but not fear) suggest additional paths may be needed to assess relatively acute GIS sensitivity. One possibility, here, is to generate scales by challenging items with anxiolytic drugs (i.e., in this case, benzodiazepines). Test a battery of items in an ABBA drug design. Those that consistently show reduced scores over the duplicate tests become your BIS/GIS scale – with score value inverted to achieve a rating of goal conflict sensitivity (which is reduced by the drugs). Then, of course, you take advantage of the linguistic nature of the items to separate state from trait. However, practical considerations are considerable with this pharmacological approach.

7. Mapping lexical Big 5 to lexical RST

Given what we have said so far, it may seem a rather large leap from the lexical/descriptive to the anchored/rational approach. We can facilitate this by passing via lexical/rational RST. Certainly, "studying individual differences requires methods distinct from studying typical function" (DeYoung et al., 2022, p. 2). But we can view the different approaches as different methods of obtaining data about the same fundamental systems – and so necessarily mapping to each other in some way. Meanwhile we must keep in mind that states represent current activity of a system; while traits represent general *input sensitivity* that then impacts on the size and neural location of activity changes. In all this, we assume, as above, that many of "those individual differences that are of most significance in the daily transactions of persons with each other" have fundamental biological explanations when those differences relate to motivation. Here we turn to Cybernetic Big 5 theory (CB5T).

This starts with lexical Big 5. But then proceeds to say

"an adequate theory of personality must explain not only how individuals differ from each other in their persisting patterns of emotion, motivation, cognition, and behavior, but also why. … A complete mechanistic theory of personality should encompass the biological basis of the mechanisms responsible for personality, and CB5T is designed to be fully compatible with the current state of personality neuroscience. … The fundamental premise of CB5T is that any adequate theory of personality must be based in cybernetics, the study of goal-directed, self-regulating systems."

DeYoung (2015, p. 33)

A key feature of CB5T, as a means of linking the Big 5 to RST, is its fundamental "goal-directed" assumption – coupled with its belief that personality should be anchored in biology. Note that, in terms of Fig. 2, DeYoung is starting with lexical Big 5 (at the social/behavioural level) and attempting to move down to the cognitive level ("goals") and then take this even further down to the neural level.The original neuropsychology, on which RST is based, is of state goal[6] control (attraction/repulsion) systems – fundamentally based in the neural level but requiring the cognitive level if it is to be applied to behaviour. To translate from the top-down trait perspective of CB5T to the bottom-up state perspective of RST requires only a mapping of the relevant goal-control (trait sensitivity, state execution) through its implicated neural systems. Goal control operating at the cognitive level is crucial for both approaches. That is, traits should be viewed as *sensitivities* of (sometimes large or multiple) systems where states can be understood as resultants in moment-to-moment action terms. Obviously, the *pattern* of state-related actions at the social/behavioural level should betray the underlying trait sensitivities when we compare individuals.

Fig. 3 shows the basic correlational relationships on which CB5T is based. We have annotated the metatraits and traits with some of the neural correlates (grey boxes) that have been proposed (see, e.g., Allen and DeYoung, 2017; DeYoung et al., 2021) – these are discussed in detail in the next section. When considering such annotation "an important caveat is that such research does not transfer directly into knowledge of causal mechanisms. … Such correlational findings can support hypotheses derived from mechanistic causal theories, but they do not provide direct evidence of causal pathways" (DeYoung et al., 2021, p. 194–5). Mechanistic state RST starts with such direct evidence. But, conversely, because of the levels of explanation problem, our mechanistic RST can have trouble seeing the (personality) wood when we are surrounded by the all-encompassing (neural state) trees. So, we have annotated the aspects and a few nominal facets with RST labels, using an essentially lexical/rational approach leaving rational anchoring for later.

Our logic, with this RST mapping, is to begin to get a picture (mediated via lexical/rational trait RST) of the *likely* mapping of lexical/descriptive trait Big 5 to anchored/rational trait RST as a guide to the final explanatory step (next section). The labelling of the Big 5 metatraits, traits, and aspects is not immediately comprehensible to an average neurobiologist. Luckily, DeYoung (2015) provides a translation of the labels in terms that are lexically very similar to those used by neural state RST. *Volatility* is defined as "Active defense to avoid or eliminate threats" and, in contrast to passive defense, this is the function attributed to the GRS by state RST. Similarly, *Assertiveness* is defined as "Incentive reward sensitivity: drive toward goals", which is the main function attributed to the GAS by state RST. Note that the traits *Neuroticism* and *Extraversion* include *Volatility* and *Assertiveness*, respectively, together with the matching aspects of *Withdrawal* and *Enthusiasm*. *Withdrawal* is annotated as including anxiety and depression. Given the GIS is defined via anxiolytic drugs that are not antidepressants, we have speculatively allowed for the effects of these two via distinct facets of goal conflict and depression, respectively. But we have also linked depression as potentially correlated to the GIS as it is the GIS that releases CORT, high levels of which generate depression (see also next section). Finally, we have added, as a facet, whatever trait is measured by the Carver and White (1994) "BIS" scale. Given its correlates this appears to match several "anxiety/worry" scales and to be linked positively to *Neuroticism* and negatively to *Extraversion* (i.e. positively to introversion, given that the *Extraversion* scale is bipolar).

A final point to note is that, in our view, RST should map only to motivation-related components of the Big 5. The RST state theory is not

[5] CORT = cortisol in humans; corticosterone in rats

[6] Note that the RST definition of a goal, e.g., McNaughton, N., DeYoung, C. G., Corr, P.J., 2016. Approach/avoidance, in: Absher, J.R., Cloutier, J. (Eds.), Neuroimaging personality, social cognition and character. Elsevier, San Diego, pp. 25–49. is more restricted than the CB5T view but is included within it.

designed to provide a complete theory of the mind nor of the Big 5, far less individual differences that go beyond basic traits. The scope of RST is shown in Fig. 3 with a background stipple. Note that this includes Neuroticism and Extraversion as these operate selectively on RST systems as higher order factors that control more than one but not all of the three basic RST systems (see next section) despite not being specific sensitivities of any particular RST system. Here, we are not saying that RST systems will not have some correlational relationships with the other three factors of the Big-5. Speculative links are possible to make, especially indirect ones (as between GIS and depression). But at present, any such relationships are not as apparent as those for Extraversion and Neuroticism.

For simplicity (and to follow DeYoung's original), Fig. 3 shows neuroticism as a sensitivity of input only to the GRS but there is good reason to see it as additionally, and equally directly, affecting processing of conflict by the GIS (as shown in Fig. 4; and discussed in the section "Understanding neuroticism"). Likewise, the figure shows the metatrait of stability as negatively correlated with the trait of neuroticism but not linked to deeper aspects or facets.

To compensate for this over-simplification, aspect-level correlation (Hovhannisyan and Vervaeke, 2021) is indicated by $\mathcal{S}$, see also "Understanding stability". There is a similar issue with plasticity, indicated by $\mathcal{P}$, see "Understanding extraversion". The full *neural trait* model, includes these different elements in the causal chain (as well as several not shown) – accounting for the presence and partial nature of inter-correlations between trait measures of stability, neuroticism, anxiety, fear, etc.; and of plasticity, extraversion, reward, etc. Thus, GRS and GIS outputs both depend on neuroticism (as well as their own individual sensitivities); while, similarly, GAS output depends on extraversion.

Now we have summarised the main elements of the revised RST and discussed the lexical/descriptive traits of personality generally, we will add detail to generate a new general theory of anchored/rational RST traits.

8. Mapping state to anchored/rational trait with RST

8.1. The basic RST traits

With the trait implications of state RST, a key fact to be born in mind is that, within RST, traits reflect sensitivities of large-scale neural systems (RST is a *sensitivity* theory). In contrast, states reflect moment-to-moment activity in those systems. As noted above, the biological factor controlling a trait sensitivity must be separate from the state system itself so (as a diffuse neural or hormonal factor) it can act on the entire detailed hierarchical neural system that controls the associated states.

The fundamental neuropsychological architecture of the RST state systems is detailed in McNaughton and Gray (2024) and can, for trait purposes, be reduced to the block-diagram shown in Fig. 4. To a first approximation, this high-level diagrammatic approach is all that matters for personality theory – which deals with *whole-system sensitivities*. The detailed neurology of each system is required only when we want to explain how a specific trait change (e.g. increased fearfulness) is converted into changes in situation-specific state behaviours (e.g., flight replacing escape or escape replacing avoidance in any particular situation when threat increases).

An important point is that RST focuses on simple goal control systems. It is not intended to explain other aspects of the mind, and hence of personality, nor to explain habitual or otherwise simple actions that are not under goal control (see, e.g., Neo et al., 2011 for neural separation of action versus goal control of stopping). As shown in Fig. 4, simple motivational stimuli (left side) provide input to goal-processing systems controlling repulsion (GRS) and attraction (GAS), each with their own sensitivities (red, green – controlling tendencies to withdraw from a goal or approach it, respectively). If the GRS and GAS are *substantially and approximately equally* activated, their inputs to the GIS activate its conflict detectors. This is related to the original idea of RST that the BIS (now renamed GIS) reflects the relative strength of introverted-neuroticism, except that it is at the aspect level while they are at the trait level (see previous section).

Output from the GIS then blocks approach and avoidance *behaviour* but leaves the *motivation* aspects of attraction-repulsion conflict intact (particularly in determining arousal). The GIS then replaces the blocked behaviours with a range of behaviours that should resolve the goal conflict. (The most salient and important of these is risk assessment.) The defensive distance ***perceived by the individual*** determines the specific conflict-resolving behaviour chosen.

Defensive distance here is an important concept. It derives, ethoexperimentally, from physical distance of a predator (cat for rat, snake for human). For any particular individual, being close to the predator reliably results in extreme active defensive reactions (fight, flight, freeze). Being further away results in less extreme, more deliberative, ones (risk assessment, exploration). However defensive distance is an internal individual construct. A more fearful individual will react to any particular physical distance as if they are closer to threat than will a relatively unfearful one. One component of this is individual differences in the sensitivity of GRS (red in Fig. 4) to its own specific inputs. The source of this specific sensitivity is not currently known[7] but acts in parallel with the more general sensitivity of neuroticism (see below).

Importantly, when there is goal conflict this decreases defensive distance relative to when there is no conflict for all individuals. But the GIS has its own specific sensitivity (amber in Fig. 4) and it is this added conflict-generated sensitivity that is reduced by anxiolytic drugs (which do not block simple avoidance in the absence of conflict). Critically, this gives us reason to see GIS sensitivity as a trait normally controlled by endogenous benzodiazepine receptor agonist ligands. In what way specific GIS sensitivity (trait GIS in Fig. 4, see also "Understanding Goal Inhibition Sensitivity") operates as a trait, given the complexity of the neural state system, is covered in the next section. Specific input sensitivity of the GAS (green in Fig. 4) operates similarly to that of GRS. Its precise basis is, similarly, not known but acts in parallel with the more general sensitivity of extraversion (see "Understanding extraversion").

8.2. The emergence of personality traits from neural states

To some extent, with RST traits, we have simple linear causal systems. The sensitivity of anything is the amount it changes for any fixed input. Higher sensitivity = bigger output per unit input. It is input that *causes* output; but, for different individuals with higher or lower sensitivity of the causal system you get more or less output (respectively) for a given fixed input. Sensitivity in this sense could vary swiftly (with addition of certain drugs, perhaps) but what we are talking about as traits are long term individual specific sensitivities. Critically a *state* sensitivity cannot be stable by definition; a stable sensitivity must be a *trait*.

We can represent such situations with $Y = a + bX$, where Y is output, X is input, a is the Y intercept at $X = 0$, and b is sensitivity. The concept of defensive distance is crucial here. For any individual, particular behaviours appear at particular physical distances (variate X in the equation) from a predator (or other source of threat). The sequence is reliable but the physical distance for a particular behaviour varies with the individual. So, defensive distance (threat perception) is an internal state construct (variate Y in the equation) and its sensitivity (parameter b) is specific to an individual. Defensive distance relates to physical distance but is modified by individual (trait) threat sensitivity ("fearfulness") to produce internal defensive distance. Individuals differ on, and drugs affect, the trait component (b). The drugs do not control behaviours

[7] Part of the control of the facet of specific panic-proneness is mediated by CCK (Wang et al., 1998). Possible association of a cholecystokinin promotor polymorphism (CCK-36CT) with panic disorder. Am. J. Med. Genet. 81, 228–234.

directly but determine where (in the real distance dimension) a specific behaviour will occur for that individual (i.e., closer to the predator with an anxiolytic).

This is where the state theory needs to step in to say that a person is experiencing Y level of fear does not, in and of itself, tell you how they will behave. Nor, importantly, how in specific terms they might score on what sort of items. With the innate responses of a rat to a cat we have ethoexperimental data that allow detailed predictions. But even with rats, learned and social responses will follow somewhat different rules. The patterns of behaviour that we see expressing any particular trait (over varying situations) are likely, then, to depend on fundamental causal chains where the final overall observed pattern emerges (like the starling murmuration) from interactions between the individual and their current physical and social environment.

8.3. Traits beyond basic RST

Superimposed on all the basic RST traits and aspects is a range of higher order metatraits and traits. We will proceed from the factor with the most widespread neural targets (stability) to factors with more restricted target (extraversion, neuroticism). Finally, we will look in detail at the operation of the most restricted fact Goal Inhibition Sensitivity. As we will see, even this most local factor, limited in its effects to a single specific RST systems, can involve not only widespread hormonal effects but also specific diffuse neural effects. Importantly the more widespread factors tend to include systems specific to the smaller scale ones.

8.4. Understanding stability

At the most diffuse trait and neural level we have stability. The trait side is fairly clear. Stability is a metatrait (including partial non-unique correlations from several traits); as is plasticity. That said, while they are each most simply viewed as intercorrelations of traits, they do have aspect influences (Hovhannisyan and Vervaeke, 2021). So, we are not dealing with simple hierarchical correlations but rather multilevel interactions (as would be expected from our proposed neurology).

At the lexical/descriptive Big 5 level, stability is thought to be controlled by serotonergic input (DeYoung et al., 2002, note the cross-over to neural allocation). Serotonin is thought to shift systems (motivationally positive as well as negative) from low level fast to high level slow thinking (Carver et al., 2008; Kahneman, 2011). Critically, subcortical serotonin is, anatomically, a pair of direct diffuse inputs that send vastly collateralised terminals to the bulk of the diencephalon (including *all* three RST systems). Serotonin is a modulator – not an element of cognitive processing – and can to some extent be thought of as a neurally released hormone. Indeed, about 90% of our serotonin is made in the gut and sends signals through the vagus nerve to brain regions including the serotonergic dorsal raphe, which can then be seen as a means of high-speed distribution of the signal to much of the forebrain (Hwang and Oh, 2025).

Serotonin, therefore, has the diffuse characteristics required of a trait control factor. Here we should flag that such a factor may not identify solely with a single compound like serotonin. Functionally, serotonin release correlates with release of noradrenaline, and to some extent dopamine. Less generally its release can match that of corticosterone. That is, a trait is a correlational object; and, where a trait reflects achievement of an evolutionary function it may have embedded in it a range of co-evolved, and so correlated, neural and hormonal components.

8.5. Understanding extraversion

To some extent likewise, there is widely distributed, diffuse, dopaminergic input to much of the forebrain. This can be linked to higher order plasticity (DeYoung et al., 2002). But there are four quite distinct dopamine systems, with different targets and less diffuse, and so more complex, neural organisation than the serotonin systems. These are the mesocortical, mesolimbic, nigrostriatal, and tuberoinfundibular pathways. They are primarily involved in cognitive processing, reinforcement, motor control, and hormonal secretion, respectively.

Dopamine can be linked separately to extraversion (Wacker and Smillie, 2015) and processing of reward prediction error, i.e. the occurrence of an *unexpected* reward (Schultz, 2016). The core of general assessment of positive valence is a dopamine-sensitive circuit involving striatum, amygdala, and orbital frontal cortex, with the latter particularly engaged in social processing (Schultz, 2015). Pure consummatory pleasure appears separate (Treadway and Zald, 2011) and likely linked to opiate mechanisms (Zurita et al., 1996). “This distinction has been described in terms of ‘wanting’ versus ‘liking’. Whereas the dopaminergic system is responsible for wanting, the opiate system is responsible for liking. … CB5T posits that the two aspects of Extraversion, Assertiveness and Enthusiasm, reflect the difference between wanting and liking [, respectively]” (Allen and DeYoung, 2017, p. 17). The main neural substrate (on which dopamine would act) of extraverted behaviour appears to include the orbital frontal cortex, anterior cingulate cortex, amygdala, and striatum (Allen and DeYoung, 2017).

We can say little more here than there is reason to see the overall control of extraversion as fitting our neurohormonal model. While there is some reason to see dopamine as a key part of the control of extraversion, there are many missing details and the precise mechanisms at the trait level are very unclear (with, e.g. reward prediction error signalling being very much a state process).

8.6. Understanding neuroticism

Neuroticism is known to impact on the processing of both fear and anxiety. It can be viewed as a sensitivity to threats generally. The Cybernetic Big 5 translation is “Defensive responses to uncertainty, threat, and punishment” (DeYoung, 2015, p. 42); with its most agreed correlate being activity in the extended amygdala (i.e. the amygdala as a whole plus the bed nucleus of the stria terminalis, DeYoung et al., 2022). Within state RST (Fig. 4) negative stimuli (NegR+, PosR-) are processed in the central amygdala to control GRS processing and in the basolateral amygdala and the bed nucleus of the stria terminalis (BNST) for GIS processing. These areas respond to certain and uncertain threat, respectively (Hur et al., 2020) and so there is reason to see them, as a whole, as where threat is processed.

To give one example, humans have an innate fear of snakes. There is a dedicated snake detection circuit that operates via a hierarchy of visual areas ranging from fast simple to slower more complex processing. All these areas send input independently to the amygdala. The early stages are responsible for the *detection* of the specific threat and for general alerting (via locus coeruleus). The amygdala is responsible for the assessment of the *level* of threat (Kawai, 2019, p. 50). The sensitivity of the extended amygdala as a whole can then account for neuroticism, which has also been called “dispositional negativity” (Hur et al., 2019).

Note that output from the BNST to the paraventricular nucleus of the hypothalamus impacts on CORT levels (purple in Fig. 4) with excessive, particularly chronic, CORT levels resulting in depression. There are a number of reports of a positive correlation of neuroticism with CORT levels (DeYoung et al., 2021, p. 200) but also some mixed results (Ormel et al., 2013). A key point is that it is only the GIS (within the RST systems) that releases CORT, via the BNST, and so any effects of GRS being captured by the neuroticism measure would dilute the neuroticism-CORT correlation. Also, there are many other areas that send input to PVN to elicit the stress response and so release CORT (again diluting the correlation with neuroticism).

The causally-serial sensitivities in Fig. 4 have implications for disorder. Neuroticism (as a general threat/negative affect sensitivity factor) is, unsurprisingly, a *risk factor* (McFarlane, 1989; Ogle et al., 2017) increasing the probability of occurrence for neurotic disorders of all

types, including depression (Andrews et al., 1990). It is also a maintaining factor, contributing to the disorder once it has occurred. But, in and of itself, it does not constitute disorder. Likewise, different neurotic disorders appear to have different, specific, pharmacological sensitivities. But all are sensitive to ketamine. It appears that each disorder requires a double hit (Fig. 5): high neuroticism (sensitive to ketamine) plus a high trait sensitivity specific to that disorder. In all cases, we can assume that sufficiently extreme events can shift trait sensitivities, generating disorder (Jordan et al., 2019; Mann et al., 2025). Note that, here, "trait anxiety" within Fig. 5 is the same as goal conflict (GIS) sensitivity. It is defined via anxiolytic drugs and does not equate with the more general use of "anxiety" in clinical practice (which would include panic, obsession, and some depression).

Precisely how this extended amygdala sensitivity is controlled is not yet known. However, as a first point, it is *not* via serotonin (this has slow parallel effects but impacts on all the RST systems, not just the extended amygdala). At present the most likely basis for neuroticism is the general control of the sensitivity of the extended amygdala (and other structures) by some endogenous equivalent of ketamine (McNaughton and Glue, 2020).

8.7. Understanding goal inhibition sensitivity

We must now drill down a level and ask how "trait anxiety" in the sense of GIS sensitivity is controlled. Unfortunately, we do not know enough to do the same for GAS and GRS. Fig. 6 shows details of the GIS, including modulating input and generation of output. You should ignore the details of the complex neurology of the goal decision and arousal systems. These are presented simply to show that explicit sets of behaviours can be predicted and are generated by a neural hierarchy (with some behaviours affected by emergence). The neural hierarchy maps to defensive distance (quick and dirty versus slow and sophisticated). Critically, choice is not determined externally. It *emerges* from interaction between the modules layered within the system (with any active higher layer inhibiting output from a lower layer – red dashed arrows).

For personality, only the left-hand side of Fig. 6 matters. All levels of the GIS (decision and arousal) receive "theta" coded input. A stronger arousal signal from the reticular formation produces higher frequency theta and this is reduced by all known specifically-anxiolytic drugs (McNaughton et al., 2007). A degree of functional purity should be obtainable via, for example, measurements of hippocampal theta activity (Fig. 6). But this still retains the state + trait problem, as does human neural functional imaging (EEG, fMRI) generally. On the other hand, structural imaging (MRI) seems necessarily trait related.

Importantly, benzodiazepines (BDZ) alter the long-term *sensitivity* of the arousal-frequency function of theta rhythm. The key to this is their mode of action on the inhibitory $GABA_A$ receptor – the benzodiazepines do not alter inhibition directly. BDZ agonists (such as Librium and Diazepam) increase GABA *site sensitivity* and inverse agonists decrease it. This change in site conformation only has effects when it is bound by GABA, producing anxiolytic and anxiogenic results respectively. Critically– they alter the longer-term sensitivity of the receptor in a truly trait-like manner. On the other hand, barbiturates, for example, increase inhibition directly by having GABA-like effects on the channel and so are far more lethal. From this point of view a person regularly taking a benzodiazepine anxiolytic is reducing their trait GIS sensitivity and there is a high probability that trait GIS sensitivity is normally controlled by endogenous ligands of the benzodiazepine receptor.

9. Chains of sensitivities

The RST systems process motivation-related inputs through a chain of neural processing elements (Fig. 4) that start with stimulus detection. We have already discussed stability, extraversion, and neuroticism as factors that directly affect multiple RST state systems. However, there are additional sensitivities that impact on the early stages of processing (Fig. 7).

The initial processing of motivational stimuli will have stimulus-specific trait-like elements (e.g. a liking for chocolate; or an aversion to spiders). These can be thought of as individual-item exchange rates. Once converted to value in a common internal currency (without which comparisons could not be made) there is an extra step determined by whether the item is being gained or lost. Importantly, experiments using dollars (which clearly have the same exchange rate as each other) have shown that $1 lost is valued more than $1 gained – a phenomenon termed loss aversion (Chen et al., 2020; Tversky and Kahneman, 1991). It is also at this stage that neuroticism is likely to have effects on threat evaluation.

These input-level changes will impact on the internal values that can then activate the attraction and repulsion systems, with further modification of the effects by those systems' sensitivities. There will be further modifications of the behavioural outputs when attraction and repulsion are balanced and activate the goal conflict system (for details of this see the previous section).

10. Output production

The final stage, once a decision has emerged from the system (approach, avoid, risk assess) is to determine the specific behaviour to be produced (Fig. 8A) and its intensity (Fig. 8B). This is where the internal motivational distance (the level of repulsion, e.g. fear; or attraction, e.g. appetite) determines which level of the relevant system is most activated. Higher levels tend to inhibit lower ones (downward negative arrows in Fig. 8A) blocking quick and dirty responses and replacing them with slow and sophisticated ones.

Note that, while specific behaviours may depend on people's training and the precise details of the current situation, this model suggests that one could assess the intensity of their *state* reactions by monitoring where in the brain is most active. Because of state-trait confounds in such experiments, as we noted earlier, more direct assessment of trait sensitivities may be obtainable with MRI (Espinoza Oyarce et al., 2022; Espinoza Oyarce et al., 2024). But note that traits based on synaptic sensitivities may not result in volumetric changes in brain structures (although as with PTSD and the hippocampus it's clear that even psychogenic trait changes can impact brain structure, Villarreal et al., 2002) and that with stress generally we are likely to see opposite changes in the amygdala, with increased reactions, and hippocampus, with decreased ones (Nadel and Jacobs, 1998).

11. Traits, RST, and emergence

Each theorist can have their own view of what is meant by "a psychological trait". We take as fundamental the views of Allport (1931) that: they have more than nominal existence; their existence may be established empirically, or at least statistically; they may be viewed either in the light of the personality which contains them, or in the light of their distribution in the population at large. That is, even with purely lexically-derived traits that have no objective anchor there is some underlying entity subject to neurobiological reductive explanation.

Reductive explanation, for us, is not completely reductive in the sense of fully explicable from the lowest levels upwards. There are multiple levels (quantum, molecular, genetic, neural, neuromorphic, and cognitive) that must all fit into any reductive explanation of behaviour – individual, social, or ecological. Critically, any level can have emergent properties (our example was of a murmuration of starlings) that will be consistent with, and explainable by, but *not* predicted directly by elements at the lower level. Finally, superimposed on these current mechanistic levels of explanation is evolutionary explanation in the form of teleonomy. This explains (in a weak sense) the current genetic level in terms of its capacity to support the intervening levels that result in behaviour that, at least in the critical parts of the evolutionary past, were adaptive and so selected. At present, however, genetic

analysis of traits does not yet seem to be on a firm footing (Montag et al., 2020). For a complete explanation, all these levels (and teleonomy) must be included.

The highest level of trait analysis is purely lexical/descriptive. The example we looked at ("Conscientiousness") can be seen as a necessary underlying mental property that gives rise to a coherent packet of verbal qualities. "That a set of characteristics load together on a factor [tells us] something that was not necessarily predicted prior to the discovery of the factor" (Ashton and Lee, 2005, p. 14). In principle, this should allow identification of the something in functional terms; prediction in advance of new correlated characteristics; and ultimately determination of the neurobiological basis of the trait. However, even with an extensively-developed and partially-anchored trait measure (Eysenck's EPQ-R neuroticism) it was clear that full functional coherence had not been determined and that the neurobiological basis was unclear. Indeed, current neural analysis of loosely defined "neuroticism" may ultimately provide the lodestone for deriving a coherent, understood, generally accepted neuroticism scale.

We then looked at the lexical/rational approach. Here the experimenters started with nominally clear and objective mental properties (autonomic perception, behavioural inhibition) and constructed scales based on their interpretation of the concepts that were then refined lexically. For various reasons, in both cases, the resultant scale (after considerable use) proved not to measure the original target construct, with verbal misinterpretation of one type or another being the prime cause.

Finally, we looked at the anchored/rational RST approach, which starts with lower-level neural state concepts and derives a view of personality traits as long-term input sensitivities of global state motivational systems. We could offer no examples of scales that have yet been appropriately developed and validated. However, it appears likely that the set of "revised RST" scales could be combined and refined (using methods similar to Eysenck's) and some already have preliminary post-hoc neurobiological tests of validation (genetic and MRI) that have promise.

In the second part of the paper, we reviewed what is known about neural state and trait systems controlling motivation. The nature of the state systems is clear (McNaughton and Gray, 2024). The trait control of the GIS appears well worked out at the neural level (Fig. 6) and allocation of potential neural control systems to metatraits, traits and aspect of the Big 5 is progressing (Fig. 3). Importantly, suggestions as to neural control of Big 5 entities can be immediately included into the theoretical neural RST model (Fig. 4) allowing explicit predictions and testing.

So, our explanations of psychological traits such as Stability, Extraversion, and Neuroticism are each part-way through a bootstrapping procedure between higher and lower levels. We have a fairly clear behaviour-level psychometric description of their every-day-use linguistic structure; reasonably well-developed suggestions as to their cognitive nature (based on the semantic coherence of their items); and preliminary global suggestions (without any high-level neural or neuromorphic detail, Fig. 2) as to their biological basis (based on known cognitive properties of well-studied large-scale neural systems). That said this linking will not be a matter of simply identifying neural bases. If we are to see serotonin as the primary trait sensitivity factor for "stability", the behavioural data on functions of the serotonin system are likely to require subtle changes in the fine-grain detail, and possibly the naming, of the concept of stability (Carver et al., 2008). Likewise, if we see "dopamine" as the primary trait sensitivity factor for extraversion (Depue and Collins, 1999; Wacker and Smillie, 2015) then we need to determine which one (or more) of the distinct dopamine systems, with their different psychological properties, underlies extraversion and also allow for involvement of opiates and oxytocin (Wacker and Smillie, 2015) – and how they relate to plasticity. This is not too dramatic a requirement as traits such as extraversion are already seen as including facets and "some facets and aspects have associations … with factors in other domains. This is true even between some traits located under different metatraits" (DeYoung, 2015, p. 36).

12. Conclusions

The 3rd edition of *The Neuropsychology of Anxiety* (McNaughton and Gray, 2024) provides a comprehensive up-to-date picture of parallel, hierarchical, goal control systems: the GAS, GRS and GIS. As we have noted, it does not provide much on trait control (with the exception of the GIS). As a neural theory it has difficulty in application to personality work. Its measures are expensive to obtain and, with the exception of structural MRI, will confound state with trait variance.

The solution to the problem of measuring the relevant traits would appear to lie in questionnaire measures (where very minor changes in wording can produce state and trait forms of the same instrument). For these we need to take a boot-strapping approach. We should use the full data (behavioural and neural) of state RST to phrase appropriate items for questionnaires – phrasing questions in terms that a rat could potentially answer (see Blanchard et al., 2001 for a test of parallel human and rat defensive behaviours), avoiding complex (particularly culture-specific) scenarios (see examples from Wilson et al., 1990 in The anchored/rational RST approach, above). It may be possible to take a criterion analysis approach (Eysenck, 1950) using a neurobiological criterion variable at the scale development stage. Another possibility, as appropriate drugs are identified, would be to test scale items for drug sensitivity. That is, items sensitive to a benzodiazepine (like triazolam), pregabalin, and buspirone (a 5HT1A agonist) would be appropriate for a trait BIS scale and those sensitive to an SSRI but not an anxiolytic could be appropriate for a trait GRS scale. For this kind of development, it might be a good idea to do the development with "state" phrasing and the convert it to trait phrasing. MRI could provide a final validation check.

In all of this we must remain sensitive to the trap of linguistic categories. The categorical nature of the reported trait of eye colour (and the resultant specific colour identities) does not reflect the underlying biological substrate. The apparently distinct colours (blue, green, brown, and several more) are subjective categories of the objective variable of black/brown and red/yellow melanin concentration (hazel involves additional light scattering). We view surface lexical/descriptive personality traits as typically measured by questionnaire in the same light: the underlying trait *sensitivities* described above seek their expression in specific lexical and cultural ways that have *psychological* meaning to the individual and to society (and to the lexicographers who assist the psychologist in uncovering semantic structure and regularities via such means as factor analysis). The lexical/rational approach will depend less on the pure use of language. In this sense, the *emergence* of personality is much like the murmuration of starlings: a sight to behold, but one that more conceals rather than reveals the causes of the underlying dynamics. It is in this very specific sense that we can say that such personality traits are not primary or causal: they are effects only, albeit highly salient, precious and interesting to each of us. They are also essential descriptions of regularities if we are then to seek causes. The causal action is at the trait *sensitivity* level, as discussed above, which then serves as the moderator of the circumstances that activate the biological systems, such as those contained in the neuropsychology of RST. Critically this level provides the mechanistic bridge between genes and the resultant behaviour, the success or failure of which can then feedback in the population to alter its gene (and behaviour) pool. This could provide a solution to the current apparent elusiveness of the genetic foundation of psychological traits (Montag et al., 2020).

We do not, then, yet have the kind of RST trait measures that we would like. But in terms of both them and other measures achieving neurobiological explanation, the future looks bright (DeYoung et al., 2022).

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.neubiorev.2026.106926.

References

Allen, T.A., DeYoung, C.G., 2017. Personality neuroscience and the Five Factor Model. In: Widiger, T.A. (Ed.), Oxford Handbook of the Five Factor Model. Oxford University Press, New York, pp. 319–352.
Allport, G.W., 1931. What is a trait of personality? J. Abnorm. Social. Psychol. 25, 368–372.
Andrews, G., Stewart, G., Morris-Yates, A., Holt, P., Henderson, S., 1990. Evidence for a general neurotic syndrome. Br. J. Psychiatry 157, 6–12.
Ashton, M.C., Lee, K., 2005. A defence of the lexical approach to the study of personality structure. Eur. J. Personal. 19, 5–24.
Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cisbani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., 2008. Interaction ruling animal collective behavior depends on topological rather than metric distance: evidence from a field study. Proc. Natl. Acad. Sci. 105, 1232–1237.
Barrós-Loscertales, A., Meseguer, V., Sanjuán, A., Belloch, V., Parcet, M.A., Torrubia, R., Avila, C., 2006. Behavioral Inhibition System activity is associated with increased amygdala and hippocampal gray matter volume: a voxel-based morphometry study. Neuroimage 33, 1011–1015.
Blanchard, D.C., Blanchard, R.J., 1988. Ethoexperimental approaches to the biology of emotion. Annu. Rev. Psychol. 39, 43–68.
Blanchard, R.J., Blanchard, D.C., 1989. Antipredator defensive behaviors in a visible burrow system. J. Comp. Psychol. 103, 70–82.
Blanchard, R.J., Blanchard, D.C., 1990. Anti-predator defense as models of animal fear and anxiety. In: Brain, P.F., Parmigiani, S., Blanchard, R.J., Mainardi, D. (Eds.), Fear and Defence. Harwood Acad.Pub, Chur, pp. 89–108.
Blanchard, D.C., Hynd, A.L., Minke, K.A., Minemoto, T., Blanchard, R.J., 2001. Human defensive behaviors to threat scenarios show parallels to fear- and anxiety-related defense patterns of non-human mammals. Neurosci. Biobehav. Rev. 25, 761–770.
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 affective responses to impending reward and punishment: the BIS/BAS scales. J. Pers. Soc. Psychol. 67, 319–333.
Chen, X., Voets, S., Jenkinson, N., Galea, J.M., 2020. Dopamine-dependent loss aversion during effort-based decision-making. J. Neurosci. 40, 661–670.
Corr, P.J., 2004. Reinforcement sensitivity theory and personality. Neurosci. Biobehav. Rev. 28, 317–332.
Corr, P.J., 2008. Cambridge. The Reinforcement Sensitivity Theory of Personality. Cambridge University Press.
Corr, P.J., Cooper, A.J., 2016. The reinforcement sensitivity theory of personality questionnaire (RST-PQ): development and validation. Psychol. Assess. 28, 1427–1440.
Corr, P.J., Mobbs, D., 2018. From Epiphenomenon to Biologically Important Phenomena. Pers. Neuro 1, e1.
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., 2006. Higher-order factors of the Big Five in a multi-informant sample. J. Pers. Soc. Psychol. 91, 1138–1151.
DeYoung, C.G., 2015. Cybernetic Big Five Theory. J. Res. Pers. 56, 33–58.
DeYoung, C.G., Beaty, R.E., Genc, E., Latzman, R.D., Passamonti, L., Servaas, M.N., Shackman, A.J., Smillie, L.D., Spreng, R.N., Viding, E., Wacker, J., 2022. Personality neuroscience: an emerging field with bright prospects. Personal. Sci. 3, e7269.
DeYoung, C.G., Grazioplene, R.G., Allen, T.A., 2021. The neurobiology of personality. In: John, O.P., Robins, R.W. (Eds.), Handbook of Personality, 4th ed. Guilford Publications, Proquest Ebook Central, pp. 193–217.
DeYoung, C.G., Peterson, J.B., Higgins, D.M., 2002. Higher-order factors of the Big Five predict conformity: Are there neuroses of health? Personal. Individ. Differ. 33, 533–552.
Draycott, S.G., Kline, P., 1995. The Big Three or the Big Five—the EPQ-R vs the NEO-PI: a research note, replication and elaboration. Personal. Individ. Differ. 18, 801–804.
Espinoza Oyarce, D.A., Burns, R., Butterworth, P., Cherbuin, N., 2021. Bridging classical and revised Reinforcement Sensitivity Theory research: a longitudinal analysis of a large population study. Front. Psychol. 12. A737117.
Espinoza Oyarce, D.A., Burns, R.A., Butterworth, P., Cherbuin, N., 2022. A new measure of the revised reinforcement sensitivity theory. Soc. Psychol. 53, 327–340.
Espinoza Oyarce, D.A., Burns, R.A., Shaw, M.E., Butterworth, P., Cherbuin, N., 2024. Neural correlates of the revised reinforcement sensitivity theory: a cross-sectional structural neuroimaging study in middle-aged adults. Psychophysiology, e14574.
Eysenck, H.J., 1944. Types of personality: a factorial study of seven hundred neurotics. J. Ment. Sci. 90, 851–861.
Eysenck, H.J., 1947. Dimensions of Personality. Routledge & Kegan Paul, London.
Eysenck, H.J., 1950. Criterion analysis: an application of the hypothetico-deductive method to factor analysis. Psychol. Rev. 57, 38–53.
Eysenck, H.J., 1957. The dynamics of anxiety and hysteria. Preger, New York.
Eysenck, H.J., 1965. Extraversion and the acquisition of eyeblink and GSR conditioned responses. Psychological Bulletin 63 (4), 258–270.
Eysenck, H.J., 1967. The Biological Basis of Personality. Charles C. Thomas, Springfield, Ill.
Eysenck, S.B., Barrett, P., Eysenck, H.J., 1984. Eysenck personality questionnaire-revised. Pergamon Press.
Eysenck, H.J., Eysenck, S.B.G., 1964. Manual of the Eysenck Personality Inventory. Hodder and Stoughton, Sevenoaks, Kent.
Fine, J.M., Hayden, B.Y., 2022. The whole prefrontal cortex is premotor cortex. Philos. Trans. R. Soc. B Biol. Sci. 377, 20200524.
Goldberg, L.R., 1981. Language and individual differences: The search for universals in personality lexicons. In: Wheeler, L. (Ed.), Review of personality and social psychology. Sage, Beverly Hills, pp. 141–165.
Gray, J.A., 1969. Sodium amobarbital and effects of frustrative nonreward. J. Comp. Physiol. Psychol. 69, 55–64.
Gray, J.A., 1970a. The psychophysiological basis of introversion-extraversion. Behav. Res. Ther. 8, 249–266.
Gray, J.A., 1970b. Sodium amobarbital, the hippocampal theta rhythm and the partial reinforcement extinction effect. Psychol. Rev. 77, 465–480.
Gray, J.A., 1971. The psychology of fear and stress. Weidenfeld and Nicolson, London.
Gray, J.A., 1976. The behavioural inhibition system: a possible substrate for anxiety. In: Feldman, M.P., Broadhurst, A.M. (Eds.), Theoretical and experimental bases of behaviour modification. Wiley, London, pp. 3–41.
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., McNaughton, N., 2000. The Neuropsychology of Anxiety: An enquiry into the functions of the septo-hippocampal system (2 ed. 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.
Hovhannisyan, G., Vervaeke, J., 2021. Enactivist Big Five Theory. Phenomenol. Cogn. Sci. 21, 341–375.
Hur, J., Smith, J.F., DeYoung, K.A., Anderson, A.S., Kuang, J., Kim, H.C., Tillman, R.M., Kuhn, M., Fox, A.S., Shackman, A.J., 2026. Anxiety and the neurobiology of temporally uncertain threat anticipation. J. Neurosci. in press, JN-RM-0704-0720.
Hur, J., Stockbridge, M.D., Fox, A.S., Shackman, A.J., 2019. Chapter 16 - Dispositional negativity, cognition, and anxiety disorders: An integrative translational neuroscience framework. In: Srinivasan, N. (Ed.), Prog. Brain Res. Elsevier, pp. 375–436.
Hwang, Y.K., Oh, J.S., 2025. Interaction of the vagus nerve and serotonin in the gut-brain axis. Int. J. Mol. Sci. 26, e1160.
Jackson, C.J., 2009. Jackson-5 scales of revised Reinforcement Sensitivity Theory (r-RST) and their application to dysfunctional real world outcomes. J. Res. Pers. 43, 556–569.
Jensen, A.R., 1958. The Maudsley personality inventory. Acta Psychol. (Amst.). 14, 314–325.
Johnson, S.L., Turner, R.J., Iwata, N., 2003. BIS/BAS levels and psychiatric disorder: an epidemiological study. J. Psychopathol. Behav. Assess. 25, 25–36.
Jordan, H.T., Osahan, S., Li, J., Stein, C.R., Friedman, S.M., Brackbill, R.M., Cone, J.E., Gwynn, C., Mok, H.K., Farfel, M.R., 2019. Persistent mental and physical health impact of exposure to the September 11, 2001 World Trade Center terrorist attacks. Environ. Health 18, 12.
Kahneman, D., 2011. Thinking, Fast and Slow. Penguin Books, Ltd, London, UK.
Kawai, N., 2019. The underlying neuronal circuits of fear learning and the Snake Detection Theory (SDT). In: Kawai, N. (Ed.), The Fear of Snakes. Springer Nature Singapore Pte Ltd, Singapore, pp. 33–58.
King, A.J., Sumpter, D.J., 2012. Murmurations. Curr. Biol. 22, R112–R114.
Krupić, D., Corr, P.J., Ručević, S., Križanić, V., Gračanin, A., 2016. Five reinforcement sensitivity theory (RST) of personality questionnaires: comparison, validity and generalization. Personal. Individ. Differ. 97, 19–24.
Leue, A., Reuter, M., Corr, P.J., Ettinger, U., 2022. Construct validity of questionnaires for the original and revised reinforcement sensitivity theory. Front. Psychol. 13. A1026894.
Mandler, G., Mandler, J.M., Uviller, E.T., 1958. Autonomic feedback: the perception of autonomic activity. J. Abnorm. Social. Psychol. 56, 367–373.
Mann, F.D., Waszczuk, M.A., Clouston, S.A.P., Feltman, S., Ruggero, C.J., Marx, B.P., Schwartz, J.E., Bromet, E.J., Luft, B.J., Kotov, R., 2025. A 20-year longitudinal cohort study of post-traumatic stress disorder in World Trade Center responders. Nat. Ment. Health 3, 789–802.
Marr, D., 1976. Early processing of visual information. Philos. Trans. R. Soc. Lond. B Biol. Sci. 275, 483–519.
Marr, D., 1982. Vis. A Comput. Approach.
McFarland, R.A., 1975. Heart rate perception and heart rate control. Psychophysiology 12, 402–405.
McFarlane, A.C., 1989. The aetiology of post-traumatic morbidity: predisposing, precipitating and perpetuating factors. Br. J. Psychiatry 154, 221–228.

McNaughton, N., 2020. Personality neuroscience and psychopathology: should we start with biology and look for neural-level factors? Pers. Neuro 3, e4.
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., 2019. Sensitivity to punishment and reward: Revisiting Gray (1970). In: Corr, P.J. (Ed.), Personality and Individual Differences: Revisiting the classic studies. Sage, Los Angeles, pp. 115–136.
McNaughton, N., DeYoung, C.G., Corr, P.J., 2016. Approach/avoidance. In: Absher, J.R., Cloutier, J. (Eds.), Neuroimaging personality, social cognition and character. Elsevier, San Diego, pp. 25–49.
McNaughton, N., Glue, P., 2020. Ketamine and neuroticism: a double-hit hypothesis of internalizing disorders. Pers. Neuro 3, e2.
McNaughton, N., Gray, J.A., 2024. The Neuropsychology of Anxiety: An enquiry into the functions of the septo-hippocampal system (3 ed. Oxford University Press, Oxford.
McNaughton, N., Kocsis, B., Hajós, M., 2007. Elicited hippocampal theta rhythm: a screen for anxiolytic and procognitive drugs through changes in hippocampal function? Behav. Pharmacol. 18, 329–346.
McNaughton, N., Smillie, L.D., 2018. Some metatheoretical principles for personality neuroscience. Pers. Neuro 1, e11.
Moncel, C., Osmont, A., Pavani, J.-B., Pichot, N., Dauvier, B., 2023. Development and validation of hierarchically structured questionnaire of approach and avoidance motivation - validation of the hierarchical reinforcement sensitivity theory of personality questionnaire. Personal. Individ. Differ. 207.
Monod, J., 1972. Chance and necessity: An essay on the natural philosophy of modern biology (A. Wainhouse, Trans.). Collins, London.
Montag, C., Ebstein, R.P., Jawinski, P., Markett, S., 2020. Molecular genetics in psychology and personality neuroscience: On candidate genes, genome wide scans, and new research strategies. Neurosci. Biobehav. Rev. 118, 163–174.
Nadel, L., Jacobs, W.J., 1998. Traumatic memory is special. Curr. Dir. Psychol. Sci. 7, 154–157.
Neo, P.S.H., Thurlow, J., McNaughton, N., 2011. Stopping, goal-conflict, trait anxiety and frontal rhythmic power in the stop-signal task. Cogn. Affect. & Behav. Neurosci. 11, 485–493.
Ng, S.H., Platow, M.J., 2024. The violent turn in non-violent collective action: What happens? Asian J. Social. Psychol. 27, 303–317.
Ogle, C.M., Siegler, I.C., Beckham, J.C., Rubin, D.C., 2017. Neuroticism increases PTSD symptom severity by amplifying the emotionality, rehearsal, and centrality of trauma memories. J. Pers. 85, 702–715.
Ormel, J., Bastiaansen, A., Riese, H., Bos, E.H., Servaas, M., Ellenbogen, M., Rosmalen, J. G.M., Aleman, A., 2013. The biological and psychological basis of neuroticism: current status and future directions. Neurosci. Biobehav. Rev. 37, 59–72.
Pandey, S., Osinsky, R., 2025. Beyond uniformity: individual sensitivities to reward and punishment shape midfrontal-theta responses to approach avoidance conflict. Soc. Cog. Affect. Neurosci. 20.
Ranganathan, P., Caduff, C., Frampton, C.M.A., 2024. Designing and validating a research questionnaire - Part 2. Perspect. Clin. Res. 15, 42–45.
Reuter, M., Cooper, A.J., Smillie, L.D., Markett, S., Montag, C., 2015. A new measure for the revised reinforcement sensitivity theory: psychometric criteria and genetic validation. Front. Syst. Neurosci. 9, 38.
Revelle, W., 2007. Experimental approaches to the study of personality. In: Robins, B., Fraley, R.C., Krueger, R.F. (Eds.), Handbook of Research Methods in Personality Psychology. Guilford Press, New York, pp. 37–61.
Rocklin, T., Revelle, W., 1981. The measurement of extroversion: a comparison of the Eysenck Personality Inventory and the Eysenck Personality Questionnaire. Br. J. Social. Psychol. 20, 279–284.
Ruch, W., Heintz, S., Gander, F., Hofmann, J., Platt, T., Proyer, R.T., 2021. The long and winding road: a comprehensive analysis of 50 years of Eysenck instruments for the assessment of personality. Personal. Individ. Differ. 169, 110070.
Sapolsky, R.M., 2004. Why zebras don't get ulcers, 3rd ed. Henry Holt and Company, New York.
Schultz, W., 2015. Neuronal reward and decision signals: from theories to data. Physiol. Rev. 95, 853–951.
Schultz, W., 2016. Dopamine reward prediction-error signalling: a two-component response. Nat. Rev. Neur 17, 183–195.
Smederevac, S., Mitrović, D., Čolović, P., Nikolašević, Z., 2014. Validation of the measure of revised reinforcement sensitivity theory constructs. J. Individ. Differ. 35, 12–21.
Smederevac, S., Sadiković, S., Čolović, P., Vučinić, N., Milutinović, A., Riemann, R., Corr, P., Prinz, M., Budimlija, Z., 2022. Quantitative behavioral genetic and molecular genetic foundations of the approach and avoidance strategies. Curr. Psychol.
Spence, J.T., Spence, K.W., 1966. The motivational components of manifest anxiety: drive and drive stimuli. In: Spielberger, C.D. (Ed.), Anxiety and behavior. Academic Press, London, pp. 291–326.
Spielberger, C.D., Gorusch, R.L., Lushene, R., Vagg, P.R., Jacobs, G.A., 1983. Manual for the STATE-TRAIT ANXIETY INVENTORY (Form Y, 94306. Consulting Psychologists Press, Palo Alto, CA.
Standen, B., Firth, J., Sumich, A., Heym, N., 2022. The neural correlates of Reinforcement Sensitivity Theory: A systematic review of the (f)MRI literature. Psychol. & Neurosci. 15, 395–422.
Taylor, J., 1953. A personality scale of manifest anxiety. J. Abnorm. Social. Psychol. 48, 285–290.
Taylor, J., 1956. Drive theory and manifest anxiety. Psychol. Bull. 53, 303–320.
Treadway, M.T., Zald, D.H., 2011. Reconsidering anhedonia in depression: Lessons from translational neuroscience. Neurosci. Biobehav. Rev. 35, 537–555.
Tversky, A., Kahneman, D., 1991. Loss aversion in riskless choice: a reference dependent model. Q. J. Econ. 106, 1039–1061.
Villarreal, G., Hamilton, D.A., Petropoulos, H., Driscoll, I., Rowland, L.M., Griego, J.A., Kodituwakku, P.W., Hart, B.L., Escalona, R., Brooks, W.M., 2002. Reduced hippocampal volume and total white matter volume in posttraumatic stress disorder. Biol. Psychiatry 52, 119–125.
Wacker, J., Smillie, L.D., 2015. Trait extraversion and dopamine function. Soc. Pers. Psych. Compass 9, 225–238.
Wang, Z.W., 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.
Watson, D., Ellickson-Larew, S., Stanton, K., Levin-Aspenson, H.F., Khoo, S., Stasik-O'Brien, S.M., Clark, L.A., 2019. Aspects of extraversion and their associations with psychopathology. J. Abnorm. Psychol. 128, 777–794.
Whitehead, W., Dresher, V., Blackwell, B., 1976. Lack of relationship between autonomic perception questionnaire scores and actual sensitivity for perceiving ones heart beat, Psychophysiology. Cambridge univ press 40 west 20th street, new york, ny 10011-4211, pp. 177-177.
Wilson, G.D., Gray, J.A., Barrett, P.T., 1990. A factor analysis of the Gray-Wilson personality questionnaire. Personal. Individ. Differ. 11 (10), 1037–1045.
Zurita, A., Murúa, S., Molina, V., 1996. An endogenous opiate mechanism seems to be involved in stress- induced anhedonia. Eur. J. Pharmacol. 299, 1–7.