7

Sensitivity to Punishment and Reward

Revisiting Gray (1970)

Neil McNaughton and Philip J. Corr

BACKGROUND TO THE STUDY

Research on personality is notoriously fragmented. Can we make it a coherent whole? In particular, can low-level (brain/body) processes explain high-level stable *patterns* of affect, behaviour and cognition, expressed in traits? Neuroscience has had a major impact on state psychology – can it help personality psychology too?

Here, we will take you 'Back to the Future'. We look at Jeffrey Gray's early sketch[1] of a neuroscientific theory of personality (Gray, 1970a), which used the concepts and experimental tools of *learning theory* (Gray, 1975) to explain the effects of extraverting drugs. Even early learning theory used mathematical models (Hull, 1943, 1952) of the *control* of behaviour. Lower-level learning constructs (habit strength, drive, goal gradients, generalization and reinforcement) mapped, implicitly, to physiological systems. Before this, Ivan Pavlov's (1927) book on 'Conditioned Reflexes' was explicitly subtitled 'an investigation of the physiological activity of the cerebral cortex'; and he viewed physiology as absolutely fundamental to an understanding of learning (Pavlov, 1932). One can, and Gray in particular did, map in both directions: between a 'conceptual nervous system' (cns) of the type inferred by Hebb (1955) from the careful observation of experimentally constrained behaviour and the real 'central nervous system' (CNS) studied by neuroanatomy and neurophysiology. Gray's unique step was to use drugs as a conceptual dissection tool – assuming that a drug changes synaptic activity (CNS) and so behaviour (cns) in parallel. Suppose a drug affects behaviour A but *not* behaviour B. We can be sure that A depends on a process not shared by B. Critically, this means we exclude from consideration all cognitive and neural processes that are not drug-sensitive. Drugs, thus, dissect *both* the cns *and* CNS in a highly replicable, *theory-independent*, way. Intracranial injections even allow microdissection of process. As we will see, we can use brain lesions similarly, whether controlled (in experimental animals) or naturally occurring (in human patients). Gray used drug dissection in a particularly powerful way: using drugs (and their parallels with specific lesions) to tie together specific behaviours, neural systems, personality systems and clinical disorders.

Gray's (1970a) cns–CNS approach has had an enormous, still increasing,[2] influence on current-day thinking – promoting the conceptual anchoring of personality traits to well-delineated brain systems. There is now even a journal, *Personality Neuroscience*, published by Cambridge University Press, dedicated to this field. But, despite this apparent progress, when you go back to Gray's classic paper you will uncover fundamental bedrock, obscured by the later rush to develop personality scales statistically. With modern neuroscience we can now incorporate fundamental principles of learning and of the real neural systems that are the substrates for all traits measured by personality questionnaires. As we shall see, Gray's overall vision in 1970 encompassed a surprising number of the types of fact that a modern neuroscientific theory of personality must accommodate but that few do. But, before visiting the past in the hope of illuminating the future, we should look at our present.

Personality research has certainly moved on from when relief was expressed that 'at least it can be said of personality that there are *now* facts and ways of gathering facts that we *can* argue about' (Claridge, 1967, p. ix, our emphasis); and it no longer has a 'Babel of concepts and scales' in which one name (e.g., 'anxiety') could include another quite different concept (e.g., 'fear'); and different names (e.g., 'emotional resilience' and 'emotional stability'), based on different scales, could refer to essentially the same trait (now, inversely, named 'neuroticism').

Indeed, 20 years ago, John and Srivastava (1995)[3] could present the 'Big Five' dimensional approach to personality traits as an emergent consensus allowing translation of the previous Babel. The Big Five dimensional axes are based on a taxonomy of patterns in natural language usage (see Chapter 5). These axes imposed order on the chaos of potential factors linked to innumerable scales; seeming to provide 'a starting place for vigorous research and theorizing that can eventually lead to an explication … in causal and dynamic terms' (John & Srivastava, 1995, p. 103).

Two decades later, we may be closer to the causal/dynamic Nirvana that they desired. For example, Cybernetic Big Five Theory (DeYoung, 2015, p. 33), based on 'the study of goal-directed, adaptive systems', provides a foundation from which the theory 'attempts to provide a comprehensive, synthetic and mechanistic explanatory model' of personality. DeYoung's approach has the potential to progress to a general-purpose model. His appeal to mechanism encourages mapping between descriptive personality traits and underlying biological causes. This echoes Gray's own theory development (Gray, 1982; Gray & Smith, 1969), which employed cybernetic principles: inputs, outputs, feedback, regulators and, particularly, comparators.

Despite these advances, the present still lacks a genuinely biological general theory of personality. Even Cybernetic Big Five Theory 'does not depend on complete or immediate translation into biological mechanisms for its utility' (DeYoung, 2015, p. 33); and so the Big Five system remains fundamentally taxonomic (i.e., it is simply a description of apparent structure and order in superficial observations).

In particular, the Big Five defines the structure of personality top-down via surface-level labels (i.e., personality traits – typically defined by self-reported patterns of affect, behaviour and cognition). But, 'taxonomy is always a contentious issue because the world does not come to us in neat little packages' (Gould, 1981, cited by John & Srivastava, 1995, p. 102). Importantly, a taxonomy based on language use – however correct in the linguistic domain – may not map to the taxonomy that ultimately emerges from lower-level causal analysis of brain processes.

We believe that the entire field of personality psychology suffers from overuse of top-down descriptions and a lack of bottom-up mechanisms. We should heed the lesson of zoology. The older top-down classification of species relationships via their superficial morphological characteristics had to be modified considerably in response to modern bottom-up genomic molecular biology (Dawkins, 2005).

A second problem all psychologists must face is inherent to any use of everyday language: the conventional meanings of our words may not map to scientific reality. In physics, an electron is a particle or a wave or neither or (paradoxically) both, depending on context and on what aspect of reality we are forcing into words. Personality may match our everyday words no better; after all, the lexicon derives from society's changing usage not biological science.

Our time travel takes us back to an important early attempt at a quite different approach to personality: where personality traits emerge, bottom-up, from the sensitivities of biological systems. Most importantly, Gray (1970a) used neural and drug data ('a reconsideration of what is known about the physiological basis of introversion', p. 257) to generate what was, in essence, a new theory of personality and almost a new approach to the entire field, albeit borrowing from early 20th-century giants such as Pavlov.[4]

The novelty of Gray's approach may not be immediately obvious since he reviews existing personality theory and learning theory before presenting the physiological conceptual bedrock that provided 'the strongest support for the present hypothesis' (p. 257). His subsequent publications and theory development were also strongly focused on pharmacology and neurophysiology (Gray, 1982; Gray & McNaughton, 2000). For Gray, descriptive personality structure (traits) always had to adjust to biological findings. Current personality science is increasingly seeking a foundation in Gray's fundamental neural, particularly pharmacological, bedrock.

Gray's (1970a) paper on 'the psychophysiological basis of introversion–extraversion' included the equally important trait of neuroticism–stability. Both traits are still very much with us as two of the current major Big Five 'domains'.[5]

Gray was inspired by the audacity of Hans Eysenck's (1967), then dominant, top-down theory of personality (see Chapter 4). Eysenck started with a medical checklist; statistically extracted a surface-level taxonomic structure of introversion–extraversion and neuroticism–stability; and, only then, searched for biological correlates (individual differences in arousability and conditionability, discussed further below). Finally, he derived a lower-level (biological) explanation of the traits from their correlates:

1. introverts have high arousal and so high *general* conditionability (i.e., a greater ease of learning, whether driven by reward or punishment), which enhances social learning and gives them an over-socialized conscience;

2. conversely, extraverts are chronically under-aroused, so learn only with difficulty, are under-socialized and prone to break societal rules.

Importantly, Eysenck's biology attempted a *causal* explanation of the differences between two types of psychiatric disorder that are both more likely in people with high levels of neuroticism. The group of what we would now call 'internalizing disorders' (e.g., anxiety and depression) generally occur in people who also score high on introversion, while the group of externalizing disorders (e.g., aggression and substance misuse but particularly for Eysenck, psychopathy) generally occur in people who also score high on extraversion.

Eysenck thus explained both types of psychiatric disorder in terms of the combination of extremes of two distinct personality traits: (a) neuroticism + introversion → internalizing; and (b) neuroticism + extraversion → externalizing. Here, neuroticism amplifies (and stability suppresses) the effects of both introversion and extraversion on behaviour, while introversion–extraversion determines the particular type of disorder that results from neuroticism.

Note that Eysenck saw the trait extremes themselves as risk factors more than disorders. Eysenck's explanation depended on a causal neuropsychological hypothesis that made the theory scientific in the sense of eminently falsifiable. We can test each step from his postulated fundamental arousal process to cognitive and social levels of explanation, particularly via his learning theory assumptions. That is, according to Gray (1970a, p. 251), Eysenck supposed the fundamental introvert property to be high general arousability, located in the brain areas controlling arousal.

People saw Eysenck's theory as problematic because it required two personality factors (introversion–extraversion *and* neuroticism) to account for both of two sets of disorders (internalizing and externalizing). One trait per set would have been neater and so Gray suggested that internalizing and externalizing could have quite separate causes: sensitivity to punishment and reward, respectively.

Gray's theory differed from Eysenck's primarily in its biology. He accepted the overall architecture and psychological/social superstructure of Eysenck's theory but used a different type of learning theory,[6] different crucial aspects of learning, and invoked different neural systems to explain introversion–extraversion. Gray's review of the literature on trait variation in arousal and sensitivity to punishment concluded that introverts condition better than extraverts *only when* there is aversive stimulation: 'High degrees of introversion represent high levels of sensitivity to punishment' (p. 259). As we discuss below, he bundled the fundamental learning theory concept of punishment together with the more esoteric one of 'frustrative nonreward'.

An important support to Gray's argument was pharmacological: non-sedative doses of alcohol and barbiturates lead to extraverted behaviour in humans but also, as is particularly well demonstrated in animal studies, only affect responses controlled by punishing stimuli and do not change responses controlled by rewarding stimuli. This *lack of effect* of extraverting drugs on rewarded learning drove Gray to conclude that extraversion does not depend on generally poorer learning and that introversion depends on a *specific* 'susceptibility to punishment', connected to a fear system. He claimed: 'The hypothesis that introversion involves a heightened susceptibility to fear (or to express the same point differently, a heightened sensitivity to punishment and warnings of punishment) has a great deal of face validity' (p. 255).

However, susceptibility to fear/punishment carried within it, like a Trojan Horse, a change in the structure of Eysenck's two dimensions of introversion–extraversion and neuroticism–stability. In particular, Gray proposed, as a corollary of his arguments about introversion, 'a new conception of neuroticism as reflecting a degree of sensitivity to both reward and punishment'. This seems not too far off Eysenck's claim of neuroticism as a general emotional *activation* factor, linked to the notion of general drive as a single process (Hull, 1943, 1952). However, Gray had split reward and punishment following the distinct two-process learning theory tradition.

Gray's (1975, p. 176) view of learning is stated as:

> essentially the same as that proposed by Mowrer in 1960 which supposes that observed learning and behaviour is the outcome of an interaction between two underlying processes: one (a classical conditioning component) responsible for the acquisition by initially neutral stimuli of reinforcing and motivational properties, the other (the instrumental component proper) responsible for the guidance of behaviour in such a way as to maximize positive reinforcement and minimize negative reinforcement.

This second step in two-process theory requires that rewarding and punishing stimuli activate distinct systems that increase or decrease, respectively, the occurrence of the behaviour that results in the stimulus. It follows that there must be two personality factors: one to reflect individual differences in reward sensitivity and one punishment sensitivity (with high neuroticism being measured when both are high). For Gray, it was important that these processes had different sensitivities to drugs and brain lesions (e.g., some drugs/lesions reduce punishment-related behaviours without decreasing reward-related behaviours, while other drugs/lesions show the opposite pattern of behavioural effects).

DETAILED DESCRIPTION OF THE STUDY

As we have noted, Gray's intellectual starting point is the biological component of Hans Eysenck's theory of introversion–extraversion and neuroticism–stability. Gray's focus on this theory and his learning theoretical approach to it are not surprising as he undertook his PhD at the Institute of Psychiatry in London, which was headed by Eysenck – who wrote an introduction to Gray's (1964c) book *Pavlov's Typology*. This 'edited' book included two chapters by Gray that were nothing short of a brilliant literal and conceptual translation of Russian psychology into Western concepts based on learning theory and arousal. There is a twist to this history: this earlier work by Gray inspired Eysenck's own 1967 arousal theory by suggesting that the 'Strength of the Nervous System' – according to Eysenck, low in introverts and high in extraverts – resulted from individual differences in cortical arousability. As another example of the strong links between Eysenck and Gray, the 1970 paper appeared in a journal, *Behaviour Research and Therapy*, founded and edited by Eysenck in 'the belief that behavioural disorders … are essentially *learned responses*, and that modern learning theory … has much to teach us' (Eysenck, 1963, p. 1). Eysenck hoped 'that this new Journal will be of interest to those who wish to apply more scientific rigour to the various fields of psychology'; and took as a motto (borrowed from the famous behaviourist John B. Watson) that 'psychology as the behaviourist views it is a purely objective, experimental branch of natural science. Its theoretical goal is the prediction and control of behaviour' (Eysenck, 1963, p. 2).

Let us briefly compare each of the levels of the distinct explanations Eysenck and Gray provide of introversion–extraversion (Figure 7.1). The primary points of difference are in the earlier levels of explanation. Eysenck saw introverts and extraverts as differing primarily in *general conditionability* (whether with reward or punishment as the reinforcer) resulting from *arousability* (shaded in Figure 7.1). In contrast, Gray suggested that they differ, instead, in *specific conditionability* (shaded in Figure 7.1), related to sensitivity to punishment (sometimes he said fear) *but not reward*. Gray's change appears very simple, but it has profound consequences for the lower levels of explanation that take us to neural systems. It also has some impact on explanations of disordered social behaviour – although, for both theories, the most important consequence of introversion for psychiatry is high conditioning of fear. (Eysenck focused on the general conditioning aspect, and Gray focused on the specific fear aspect.) For both theories, high/low fear conditioning results in high/low socialization, respectively. Both theories presumed that these introversion-/extraversion-based differences in socialization would lead to psychiatric disorder when combined with high levels of neuroticism, which acts like an amplification factor. Despite this similarity in primary social and psychiatric predictions (based on conditioning via punishment), the two theories differ in their predictions about conditioning via reward. However, Gray's approach provides a much more nuanced account of the types of behaviour, derived from learning theory, to characterize internalizing and externalizing disorders.

The theory presented in Gray's paper as a whole links arguments between these various levels of explanation. It also involves novel suggestions at each level. We will look at the elements of Gray's argument using his original section headings.

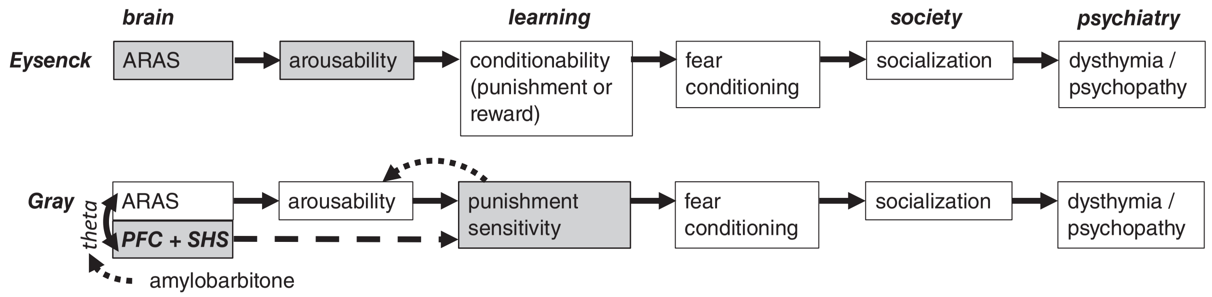

Figure 7.1 Eysenck's theory of introversion–extraversion compared with Gray's as shown in Figure 2 and Figure 6 of Gray (1970a) with additions based on his text. Both theories presume that conditioning of fear is high in introverts and low in extraverts and so, at the social and then psychiatric level, introverts with internalizing disorders (dysthymia) are over-socialized and extraverts with externalizing disorders (psychopathy) are under-socialized. Gray's key modification (shaded) is to attribute variations in fear conditioning to differing sensitivities to punishment, whereas Eysenck attributes it to variations in general arousability (in the ascending reticular activating system, ARAS) and so, as a consequence, conditionability in general. Gray located punishment sensitivity (in the sense of susceptibility to fear; see Gray, 1970a, p. 255) in the prefrontal cortex (PFC) and the septo-hippocampal system (SHS), as shown by the dashed arrow. He connected PFC, SHS and ARAS in a feedback loop, controlled by 'theta' rhythm, and impaired by extraverting (anti-anxiety) drugs such as amylobarbitone. High arousal could generate punishment (with effects similar to those proposed by Eysenck). Conversely, high punishment sensitivity would generate high arousal in punishing situations (dotted arrow) due to interaction of PFC + SHS with ARAS.

Conditionability

Much of the debate on personality in the human conditioning literature revolved around a particular type of conditioning, namely that of the eyeblink. Gray's first data-oriented section focuses on eyeblink conditioning in both introverts and those high on 'Manifest Anxiety' (Taylor, 1956), who he argues (via his Figure 3) are neurotic introverts. The eyeblink conditioning data, and arguments, are complicated (particularly where partial reinforcement schedules are used) but best fit the idea that introverts learn better than extraverts *only* under conditions where they are more highly aroused; with those high on trait anxiety (i.e., neurotic introverts) showing better conditioning when exposed to threat. In passing, Gray suggested that this trait arousability is equivalent to Pavlov's 'Strength of the Nervous System',[7] which we have already come across above derived from analysis of individual differences in conditioning (Gray, 1964a). While working through the arguments Gray presented in this section, you should bear in mind that eyeblink conditioning is aversive (see below) and that, in any case, arguments about conditioning in general would be better if based on more than one paradigm.

To understand the eyeblink conditioning paradigm, imagine yourself in Eysenck's laboratory at the Institute of Psychiatry in London. You seat yourself in a comfortable chair in a small room some 6 by 10 ft (you have time-travelled prior to metrication). White metal plates cover the walls and have holes in them to dampen reflected sound (the same as in sound recording studios of that era). On the wall directly in front of you, a small spot of red light appears. Shortly after, a device attached to your eye delivers a puff of air, which makes you blink. After a number of such trials, you will blink when the light occurs and before the air puff.

In learning theory terms, this is a classical conditioning procedure in which a (to be) *conditioned stimulus*[8] (CS; the light) is reliably and swiftly followed by the *unconditioned stimulus* (US; air puff), which elicits the *unconditioned response* (UR; eyeblink). A sensor over the eye carefully records the response producing a trace automatically recorded on paper in an adjoining room. After enough CS:US pairings, the CS *alone* is enough to trigger the eyeblink, which in the absence of the US we call the *conditioned response* (CR). Importantly, the traces of the UR and CR are somewhat different. We can score learning as the strength of the CR after a fixed number of trials, or the number of conditioning trials needed to reach some criterion strength. During 'extinction', when the CS occurs alone (i.e., it is not reinforced by the US), we can measure the number of trials needed to reach some criterion of non-response.

Neurotic introverts usually condition eyeblinks faster and extinguish them slower than other people. If we can generalize from this to all learning (particularly social), we can then account for their introverted symptoms in the same way as Eysenck. As you might well imagine, eyeblink conditioning is (mildly) unpleasant. If we assume that introversion, especially with high neuroticism, amplifies the unpleasantness we can account for the eyeblink results in the same way as Gray.

Sensitivity to punishment and nonreward

If high conditioned fear, as shown by the eyeblink-conditioning paradigm, is not due to *better conditioning* in general, Gray suggested, it could be due to *susceptibility to fear* and particularly its induction by punishment. Susceptibility to *fear* (although not always due to conditioning, see below) fitted well with a number of facts (p. 255). We can easily see internalizing disorders ('dysthymias', e.g., phobia, anxiety and obsession) as excessive fear of one form or another. As we noted, eyeblink conditioning is aversive; furthermore, trait-anxious people (neurotic introverts) condition better only if there is threat. At the other end of the scale, we can view externalizing disorders (e.g., psychopathy) as insufficient sensitivity to punishment. Of course, for proof that good conditioning is selective to fear/punishment 'the crucial test would be [of] introverts and extraverts using non-aversive reinforcement in a definitely unthreatening environment; but … no such experiment [had] yet been carried out' (Gray, 1970a, p. 255) – but this experiment did follow (e.g., Corr, Pickering, & Gray, 1995).

Depression seems to stand apart from fear, anxiety, punishment and conditioning; but, like 'other dysthymic neuroses (i.e., phobias, anxiety state and obsessive compulsive neurosis'; p. 256), it is related to introversion and high neuroticism. Gray accommodated depression, perhaps surprisingly given its nature, via his first detailed application of learning theory. His immediate problem was reactive depression resulting from loss of reward (e.g., death of a spouse), not punishment. In his solution, we can see the power of a properly formulated learning theory perspective of the kind urged by Eysenck (1963, p. 1). To understand Gray's argument, we need to take a step back. Gray's primary hypothesis concerned punishment. So, he obviously needed to equate loss of reward with punishment. Serendipitously, he had previously proposed the 'fear = frustration hypothesis' to explain emotional reactions (Gray, 1967). All Gray said in 1970 was that 'the evidence for this hypothesis is rather strong' (1970a, p. 256); but you can check this evidence in his 1967 paper and in his later book *The Psychology of Fear and Stress* (Gray, 1971, 1987). Briefly, when an animal *fails to* receive an *expected* reward its immediate reactions (increased arousal, escape, attack if a conspecific is present) are 'functionally and physiologically very similar, and perhaps identical' (1970a, p. 256) to when it receives a shock (or other punisher). The reactions show that failure of expected reward generates an emotional state, usually called 'frustration', and this has received extensive analysis (Amsel, 1992). Gray's conclusion is that introverts, who are more sensitive to fear, will also be more sensitive to frustration in the extreme form generated by severe loss, and so are more likely to become depressed.

The physiological basis of introversion – drugs

The core of Gray's argument is a new proposal for the neural substrate of introversion, which he based on a learning-theory-driven overview of the effects of extraverting drugs. He thus linked behaviour to neural systems by using extraverting drugs as a kind of tracer.

The extraverting drugs had made a major contribution to his previous paper on the 'fear = frustration hypothesis' (Gray, 1967). Crucially, their pattern of effects on conditioned and unconditioned responses is the same with frustration as it is with fear. His 1967 paper focused on ethanol and barbiturates,[9] particularly amylobarbitone – but its conclusions have proved true for modern anti-anxiety agents. 'Anxiolytic' drugs, both classical and novel, reduce *response suppression* and partial schedule effects similarly, whether we omit expected food or present shock (Gray, 1977; Gray & McNaughton, 2000, Appendix 1). *In that limited sense*, as Gray claimed, they do 'reduce the effects of punishment and of frustrative nonreward' (1970a, p. 257).

Equally important for Gray's argument was the complementary 'hope = relief hypothesis' (Gray, 1971, 1972),[10] derived from his concept of relieving nonpunishment (a mirror image of frustrative nonreward). Extraverting drugs *do not impair avoidance* unless some form of conflict is present (i.e., avoidance is passive, not active – a subtle but fundamental distinction). *Provided we are dealing with learning*, we can see an active avoidance response as one rewarded by stimuli that signal safety and generate the positive emotion of relief; and so we can explain the lack of effect of anti-punishment drugs. (Escape and related forms of active avoidance represent withdrawal from fear not approach to relief.) Results in 'the Miller–Mowrer shuttle-box' are particularly interesting (1970a, p. 257). This apparatus was popular because it automated instrumental conditioning; but you will need a bit of thought to understand the effects of extraverting drugs in it. The shuttle-box has two adjacent compartments; and we train the animal to shuttle between them through a door. The animal starts in one compartment. We present a tone followed by a shock and the animal escapes to the next compartment. After a pause, we present the tone and then shock – this time in the second compartment – and the animal shuttles back. After several repetitions, the animal will shuttle at the tone and so not get shock – this is learned avoidance. Perhaps unexpectedly, extraverting drugs *improve* shuttle-box ('2-way active') avoidance. Why? First, note that early on the animal has to escape into a compartment in which it received a shock on the immediately previous trial. The expected punishment will produce a tendency to passive avoidance that will slow escape and active avoidance. Gray explained amylobarbitone's improvement of 2-way avoidance as a reduction in conflict[11] between primary, correct, active avoidance (known to be unaffected by the drug) and secondary, interfering, passive avoidance (known to be reduced by the drug). This pattern of effects across the three avoidance paradigms will be particularly important when we compare drug and lesion effects below.

The physiological basis of introversion – brain

Eysenck identifies the Ascending Reticular Activating System (ARAS in Figure 7.1) as the substrate for arousal at the state level and arousability (hence introversion) at the trait level. At the neural level, Gray retained a contribution from the ARAS but restricted its importance by embedding it within an important feedback loop involving the prefrontal cortex and hippocampus (PFC and SHS in Figure 7.1). A key element of his neural argument was evidence that extraverting drugs (barbiturates and alcohol) impair slow rhythmical activity 'theta' that functions to coordinate both hippocampal and orbital frontal cortex function. Gray's basic argument, here, is that these drugs reduce introversion (and move the individual in the direction of extraversion) and reduce theta; and so systems that depend on theta are likely to be the substrate of introversion.

The key neural component for Gray is the septo-hippocampal system. The hippocampus is renowned for its theta rhythm – one of the strongest sine wave-like rhythms recorded from the brain. Critically, Gray had shown that amylobarbitone (the main drug of his behavioural review) impaired the control of hippocampal theta by its pacemaker in the medial septum particularly at a frequency that he had shown occurred when the animal experienced frustrative nonreward (Gray, 1970b; Gray & Ball, 1970).[12] Crucial for Gray's proposal of the septo-hippocampal system as the functional site of anxiolytic action was that septal and hippocampal lesions are like amylobarbitone in that they 'impair passive avoidance and extinction of once-rewarded behaviour, but enhance two-way active avoidance' (Gray, 1970b, p. 466). This drug–lesion parallel has held up and been massively extended over the years (Gray, 1982; Gray & McNaughton, 1983, 2000).

Gray (1970a) added two extra ingredients to this primary septo-hippocampal hypothesis of amylobarbitone action:

(a) 'the similarity between the effects of this drug in Man and the effects of damage to the frontal cortex in Man' (p. 259); and

(b) that 'lesions to the frontal cortex ... produce the same pattern [of behavioural effects as] lesions to the septal area and lesions to the hippocampus' (p. 259).

It is important to note that 'pattern' here refers not only to the dysfunctions produced by the drugs on some measures but, equally importantly, to the *lack* of dysfunction on others (which dissect away some processes) on a battery of tests. Gray scatters the parallels through his paper and so we summarize them in Table 7.1. Comparing the rows, we can conclude that the treatments have essentially the same effects as each other across the battery of paradigms. Comparing the columns, we can conclude that learning to produce an active response is unimpaired (contrary to Eysenck's theory if we believe the treatments are extraverting) while learning to suppress responding is impaired. This specificity to suppression later became the foundation for Gray's most influential suggestion: that the brain contains a distinct Behavioural Inhibition System (Gray, 1975, p. 250; 1976; see particularly Gray, 1977). From all this work, Gray concludes 'that *it is activity in this frontal cortex-medial septal area-hippocampal system which determines the degree of introversion*: the more sensitive or the more active this system is, the more introverted will the individual be' (1970a, p. 260).

Table 7.1 Summary of the effects of Gray's treatments of interest across a range of learning paradigms

Treatment	Rewarded learning	Rewarded extinction	One-way avoidance	Passive avoidance	Two-way avoidance
Anxiolytic drug	0	–	0	–	+
Septal lesion	0	–	0	–	+
Hippocampal lesion	0	–	0	–	+
Frontal lesion	0	–	0	–	+

The important point to note is that not only are the deficits the same across treatments but so are the failures to have an effect. Critically in contrasting rewarded learning with extinction and one-way active avoidance with passive avoidance we can conclude that the treatments are reducing response suppression but not response learning (which is actually improved in the case of two-way avoidance). You can view rewarded extinction as a form of passive avoidance resulting from frustrative nonreward.

Arousability and sensitivity to punishment

Gray wanted his new approach to accommodate Eysenck's existing theory as far as possible. He says, 'it would be in the interests of parsimony if we could now relate differences in susceptibility to punishment to differences in arousability *in the same way that Eysenck* relates conditionability to arousability' (1970a, p. 26, our italics). According to Gray, arousability is a general concept that should apply to both reward and punishment. To explain activity in a 'system whose chief function appears to be that of inhibiting maladaptive behaviour' (p. 260), general arousability needs explanation. For Gray, if we invert the causal order, it seems perfectly reasonable that higher susceptibility to the threats that abound in everyday life would lead to higher levels of arousal. But what about the effects of arousal highlighted by Eysenck? Gray noted that 'any stimulus, if it is made sufficiently intense, may act as a punishment'. High enough arousal will be punishing and so its performance-impairing effects are punishment-mediated – neatly explaining Eysenck's U-shape arousal-performance effects. For Gray, arousal, however it is produced, serves to invigorate behaviour (his Figure 5), unless it is so intense it becomes punishing. This can give rise to paradoxical effects: for example, mild punishment will induce arousal and may invigorate reward-mediated reactions – so long as the punishment-inducing effects are smaller than the reward-inducing effects.

Nevertheless, Gray, rather surprisingly, tries to retain Eysenck's suggestion that the ARAS is the key neural structure. He does this by saying that high ARAS activity would feed, via the medial septum, into changes in the hippocampal theta rhythm, and so changes in the hippocampus and frontal cortex. The whole point about the reticular (net-like) aspect of the ARAS is that it sends neural tentacles everywhere. As with his treatment of arousability, then, Gray ignores the general impact that the ARAS has on the brain, focusing on the frontal cortex-medial septal area-hippocampal system (and so deriving at the neural level punishment-specificity). As mentioned above, we can expect ARAS-induced arousal to augment reward-related reactions too, although at low levels of arousal we should expect this effect to be mild and, as we have already seen, with increasing intensity of arousal, punishment-mediated processes are most likely to dominate. In later work Gray focused almost entirely on the septo-hippocampal system.

The nature of neuroticism and anxiety

This final section is where the Trojan Horse of punishment sensitivity unleashes unexpected effects on the relationship of Gray's theory to Eysenck's. So far (especially in his attempt to include arousal), Gray could be seen to simply provide a slight modification to the neural elements of Eysenck's theory without much altering its superstructure (right-hand side of Figure 7.1). Moreover, in this section on neuroticism and anxiety, Gray first accepts, apparently wholeheartedly, the idea that neurotics have more intense emotional reactions both positive and negative. Neuroticism, here, is akin to emotionality. Then a twist appears.

Taking an explicitly two-process learning approach, Gray first recasts the combination of neuroticism with introversion. If reward and punishment sensitivities are distinct, and we employ only two factors for our explanations, then high neuroticism/emotionality as normally measured must represent a *combination* of high reward and high punishment sensitivity. Gray's initial equation of introversion with punishment sensitivity means that the *neurotic* introvert will be particularly sensitive to punishment. From this position, we would expect that those high on the Manifest Anxiety Scale (or suffering from any internalizing disorder) would be neurotic introverts, as his Figure 3 shows. Gray also simply asserts as a corollary ('which we now offer as a more precise statement of the present hypothesis'; 1970a, p. 262) that those with externalizing disorders would be neurotic extraverts: particularly sensitive to reward.

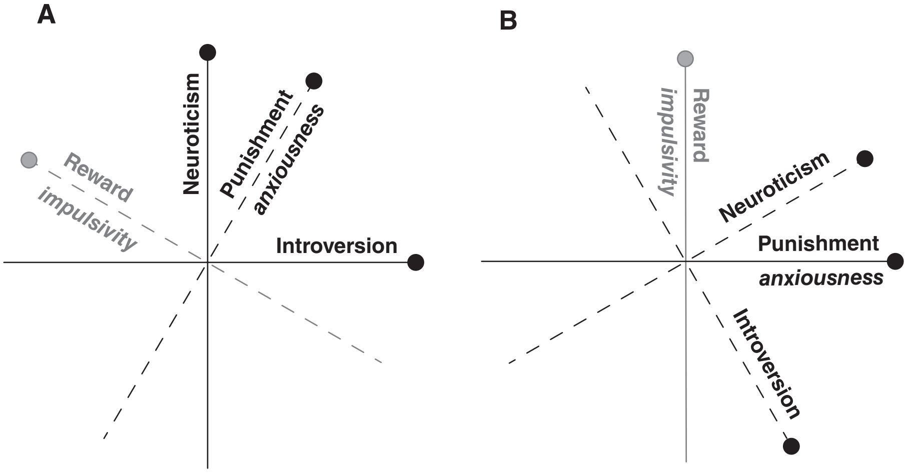

Figure 7.2 Relationship of Gray's factor axes of trait reward sensitivity (impulsivity) and trait punishment sensitivity (anxiousness) to Eysenck's factor axes of neuroticism and introversion (see also Pickering et al., 1999). (A) Treating Eysenck's factors as primary, note that the top right quadrant matches Gray's Figure 3. (B) With the space rotated to treat Gray's factors as primary.

Gray then presents in his Figure 7 a translation between susceptibility to punishment/reward and the combination of introversion and neuroticism. Here he said only that 'the most rapid increase in absolute sensitivity to punishment' (in which he included frustrative nonreward) should match the position of Taylor's Manifest Anxiety (Taylor, 1956). His footnote makes clear that this punishment sensitivity/anxiousness axis is equivalent to a mix of one-third introversion and two-thirds neuroticism as plotted in his Figure 3 (redrawn as the upper right quadrant of our Figure 7.2A). However, he also implied a second independent axis of trait sensitivity to reward (i.e., trait impulsivity; added in grey in Figure 7.2A.). In taking the two most basic concepts of learning theory (reward and punishment) as the basis of key independent traits, Gray rotated Eysenck's parameter space (Figure 7.2B) in a functionally important way (made quite explicit in Pickering et al., 1999). This rotation triggered a minor revolution, which was to go beyond Eysenck's biological theory of introversion–extraversion, and neuroticism–stability and became a completely new field. Note, here, that there are some subtle problems with Figure 7.[13] The first is that, as drawn, it implies an interaction between neuroticism and introversion, with higher neuroticism increasing values disproportionately (see Gray & McNaughton, 2000, p. 336). The second, mentioned in his footnote, is that its reward and punishment effects are symmetrical (i.e., an axis rotation of 45° not the 30° shown in Figure 7.2).

IMPACT OF THE STUDY

Gray's 1970[a] classic study had an enormous influence on the field of personality psychology. Maybe its most important achievement was to provide a different way of thinking about personality factors and their biological basis. Although Gray took his lead from Eysenck (and Pavlov) – informed by a rich literature of brain-behavioural psychology – he went further than anyone else at the time to show how a sophisticated approach to learning theory and neurophysiology, especially the use of drugs as an experimental tool, can contribute to understanding the causal dynamics of personality. This came at a price, which it still pays today: Gray's approach is complex; can be difficult to understand; and requires a breadth of not only disciplines but also technical approaches that can make it difficult to implement in personality psychology.

In consequence of Gray's work, today there is a large and growing family of approach–avoidance personality theories,[14] which include the differing approaches of Elliot and Thrash (2002, 2010), Cloninger (Cloninger, 1986; Cloninger, Svrakic, & Przybecky, 1993; Gardini, Cloninger, & Venneri, 2009), Depue (Depue & Collins, 1999; Zald & Depue, 2001), Davidson (Davidson, Ekman, Saron, Senulis, & Friesen, 1990; Davidson, Shackman, & Maxwell, 2004) and Carver (Carver, 2005; Carver & Harmon-Jones, 2009; Carver & White, 1994).

Gray's own fuller development of his ideas (Gray, 1982) has been further extended by students and former colleagues, including us (Corr & McNaughton, 2008, 2012; Gray & McNaughton, 2000; McNaughton & Corr, 2004, 2014) in what is now called the Reinforcement Sensitivity Theory (RST) of personality (Corr, 2008). RST is now a complex neuropsychological theory of personality and clinical disorder (e.g., Corr & McNaughton, 2016; McNaughton & Corr, 2016) with known links to neuroimaging data (McNaughton, DeYoung, & Corr, 2016). Other researchers around the globe have also made significant contributions to these empirical advances.

CRITIQUE OF THE STUDY

Gray's 1970[a] synthesis was unprecedented and fundamentally changed personality psychology. However, his approach had several problems, some of which linger to this day. Most obviously, its theoretical elegance is fraught with complexity. His neural and psychological *state* theory has now been greatly extended; but even this updated theory has not been easy to translate into human personality psychology; although there have been major recent moves in this direction.

The paper's complexity may seem a trivial issue – something one expects scientists to deal with. However, even half a century later, readers (including us) struggle with it. The biggest problem is that the theory spans multiple disciplines – with each integral to the whole. It is a major strength for a theory to explain more of the data: less complete theories fall before Occam's razor.[15] However, the range and depth of Gray's multidisciplinary arguments make them impenetrable for those who normally work within only one of the contributing disciplines.

Gray's detailed exposition also has some specific problems that we discuss here. At the theoretical level, his use of the terms 'punishment' and 'fear' were ambiguous: blurring key points when he shifted between one and the other conceptually. At the measurement level, while proposing a rotation of Eysenck's axes, he did not tell us how to assess his proposed reward and punishment sensitivities – a psychometric issue that troubles us to this day.

'Punishment' (or its alter ego 'frustrative nonreward') can suppress ongoing responding (passive avoidance); generate approach to safety; or elicit escape/withdrawal. Extraverting drugs reduce only the effects of the first of these three effects. For the second case, Gray argued that, for example, shuttle-box avoidance is unimpaired by the drugs because *relief* rewards avoidance rather than fear punishing non-avoidance. His paper focused on 'reward' and 'punishment' in the context of conditioning. He, therefore, did not discuss the third case of escape/withdrawal in any detail. At this time, he distinguished between a learning-related 'punishment mechanism for passive avoidance [and] a separate punishment mechanism for organizing the unconditioned response to a punishment. We shall call this the "fight/flight" system' (Gray, 1971, p. 194; note that 'punishment' means three different things within this one quote). However, it is via fight/flight that he included obsessive–compulsive disorder, with its compulsive rituals and obsessive rumination, within the dysthymic disorders. Gray says the 'symptoms bear all the marks of an *active* avoidance response' (p. 255, our italics) equating this with *fear* (see also Rapoport, 1989). However, we cannot link this (active avoidance) 'fear' to (passive avoidance) 'punishment' behaviourally; nor do extraverting drugs (like amylobarbitone) treat obsessions or compulsions. Other, similar, mismatches occur in the clinic. Neuroticism appears to be a quite general risk factor for a wide array of dysthymic disorders (Andrews, Stewart, Morris-Yates, Holt, & Henderson, 1990; Hengartner, Tyrer, Ajdacic-Gross, Angst, & Rossler, 2017). However, many of these disorders do not involve passive avoidance (not just obsessive–compulsive disorder, but also simple phobia, panic, depression) and do not respond to Gray's key pharmacological tool: the extraverting drugs[16] that reduce behavioural inhibition in animal tests and reduce generalized anxiety in humans. In explaining reactive depression, Gray notes that Maudsley Reactive rats, taken as a model of introversion, 'show a bigger frustration effect in the Amsel and Rousell (1952) double runway' (1970a, p. 256) but he does not note that he had already shown that amylobarbitone does *not reduce* the frustration effect (Gray, 1967, p. 601). Gray's 'punishment sensitivity' as defined by the drugs, even at the time, was more restricted than neurotic introversion or dysthymia.

Gray set out to replace Eysenck's theory of introversion and neuroticism with his own; but as RST has evolved, he appears more to have provided an explanation of how neuroticism and introversion give rise to the psychiatric disorders that were Eysenck's primary starting point. Given his bottom-up biological approach, it may seem surprising that Gray retained Eysenck's two-dimensional personality space, simply rotating the introversion–extraversion and neuroticism axes to form punishment and reward sensitivity factors (Figure 7.2). He retained this structure even in the substantial revision of the state theory by Gray and McNaughton (2000).

Many personality researchers have sought specific scales for Gray's biological factors (see Corr, 2016b). But even Gray's own attempt, the Gray–Wilson Personality Questionnaire (Wilson, Barrett, & Gray, 1989; Wilson, Gray, & Barrett, 1990), does not have the predicted factor structure. The most recent attempt of this kind, the Reinforcement Sensitivity Theory Personality Questionnaire (Corr & Cooper, 2016), is the most elaborate and professes to have been developed exclusively on the basis of the most recent state version of RST.

However, none of these attempts (including Gray's own) has used biological anchors as a starting point and all have assumed that the experimenter's use of language will map to the underlying biology. One reason for this is that many personality researchers prefer the persuasive mono-disciplinary simplicity of language, as reflected in the Big Five and, so, do not anchor their personality constructs in well-delineated biological systems (McNaughton & Corr, 2014). This is a preference Gray warned us against all those years ago in 1970, but the lesson has still to be learned. This warning may well be Gray's truly lasting legacy to personality psychology. A second reason is that development of Gray's state theory did not offer, until recently, an obvious anchor that personality theorists could use. However, based on the fundamental ideas in Gray's (1970b) paper on conflict and drug-sensitive theta rhythm (and some decades of practical development) we now have a human biomarker for Gray's Behavioural Inhibition System that offers the first such (albeit weak) biological anchor for personality research (McNaughton, 2017). It will be interesting to see if personality researchers wish to take it up.

CONCLUSIONS

You have just seen how Gray's 1970[a] classic study contributed to personality psychology. His theoretical brilliance is not in doubt: he posed new and exciting questions for the 'student of personality'. The paper showed that we should anchor personality measures to known biological entities; and it anticipated the links research is now forging between personality and the psychobiology of the mental illnesses that still blight the lives of millions with high costs to society. Gray was right to not simply accept that statistically derived lexical factors reflect the true nature of personality. He did accept that we need them as a starting point for biological exploration – all science has to start with a superficial descriptive phase. But, his multidisciplinary sophistication is still not a feature of personality psychology, where most research workers are yet to grapple with the more fundamental biological reality that underlies systematic individual differences in patterns of affect, behaviour, cognition and desire: personality.

NOTES

1. N. McN. remembers him saying that, like Archimedes, the crucial ideas came to him during a bath and he wrote the paper straight out in a day or two.
2. Gray (1970a) has recently had steadily increasing citations, averaging about 10/year around 2000, 20/year around 2005, 25/year around 2010, and well over 30/year around 2015 according to Web of Knowledge.
3. They provide a broad coverage of the history and development of the Big Five.
4. See Pavlov (1927), Lecture XVII, p. 284 on 'The different types of nervous system' and their links to pathology and Gray's (1964c) overview of Pavlov's theory of types and its mapping to Western views of personality and arousal.
5. The other three are Conscientiousness, Agreeableness, and Openness to Experience.
6. In the history of learning theory, there have been two main traditions: single-process (used by Eysenck, following Hull) and two-process (which we will meet shortly). In Hull's theory, all learning depends on a single process, drive reduction (Hull, 1943, 1952) – reinforcement occurs when a stimulus (e.g., water, money, or verbal praise) reduces drive. Concurrent active drives summate, and the amount of conditioning (based on drive reduction) should increase with increases in general arousal (which reflects increased summated drive), except when drive/arousal is too high. The idea of an inverted-U relationship between arousal and performance goes back at least to the time of the Yerkes–Dodson law (Yerkes & Dodson, 1908). Eysenck's insight was that introversion–extraversion could map to this arousal–performance relationship, with both extremes producing sub-optimal performance. (For a discussion of this literature, see Corr, 2016a, pp. 115–130).
7. Strictly (Gray, 1964b, p. 158) this is what Pavlov would have called the strength of the excitatory process (which would include transmarginal and external inhibition) and is distinct from his strength of the inhibitory process (involving

internal inhibition). However, like Gray (1970a), we can ignore these details in concluding that Eysenck's theory does not match the eyeblink data.

8. 'According to one story, the English versions of Pavlov's writings were first translated with the guidance of the German translation from the original Russian. The German word "bedingt" has two meanings that have different words in English: "conditioned" and "conditional". It was translated as the more common conditioned, but Conditional and Unconditional are more accurate translations of the Russian, and they fit the underlying idea of conditioning.' www.indiana.edu/~p1013447/dictionary/origcond.htm
9. A class of drug abused by, and one of the causes of death of, Marilyn Monroe.
10. Gray (1972) is the same chapter as that cited by Gray (1970a) as 'Gray, in press, b'. However, it appeared in a different book than that in the original reference, with Gray's contribution to Cattell's book being on a different topic.
11. As a cherry on this icing to his cake, Gray provided in his Figure 5a sketch of the Gray and Smith (1969) 'arousal-decision model for partial reinforcement and discrimination learning' (also reprinted in Gray, 1975). Your most important take-home message from the model is that Gray saw it as 'a model for conflict situations', by which he very much meant the kind of situations analysed behaviourally and pharmacologically by Neal Miller (Bailey & Miller, 1952; Barry & Miller, 1962, 1965; Kimble, 1961; Miller, 1944, 1959). In these situations, avoidance *opposes* approach and – depending on goal gradients and other factors – the animal may approach, avoid passively, explore, or dither. Whether approach occurs, or not, depends on the decision mechanism (the box [D.M.] in Figure 5). How fast the individual acts, and things like whether they explore (which we can view as risk assessment) or how much they dither depends on the arousal summation mechanism (the box [A] in Figure 5). Note that general effects of the functions of [A] are very similar to those of Hull's generalization of drive that we discussed earlier.
12. Gray (1970b) is nominally a review article but, unusually, contains original data. The key findings were reported in *Science* by Gray and Ball (1970), which is not cited by Gray (1970a). Gray produced all three papers during his visit to Neal Miller's laboratory.
13. In 1999, Gray acknowledged these facts in a *mea culpa* when Dr Alan Pickering pointed them out (Pickering et al., 1999).
14. For a review of influential theories in personality neuroscience, see DeYoung and Gray (2009).
15. 'Pluralitas non est ponenda sine necessitate' (keep assumptions to the minimum necessary to explain the available data) Guilelmus de Occam: Quodlibeta, V, Q.i.
16. For Gray (1970a) the key drugs were barbiturates, usually sodium amylobarbitone, and also alcohol. The same lack of effect on fear disorders has proved true of currently prescribed anxiolytic drugs, both 'classical' benzodiazepines and more novel 5HT1A agonists like buspirone and calcium channel agents such as pregabalin.

FURTHER READING

Elliot, A. J., & Thrash, T. M. (2002). Approach-avoidance motivation in personality: Approach and avoidance temperament and goals. *Journal of Personality and Social Psychology, 82*, 804–818.

Gray, J. A. (1981). A critique of Eysenck's theory of personality. In H. J. Eysenck (Ed.), *A model for personality*. New York: Springer.

Gray, J. A. (1987). *The psychology of fear and stress*. London: Cambridge University Press.

REFERENCES

Amsel, A. (1992). *Frustration theory: An analysis of dispositional learning and memory*. Cambridge: Cambridge University Press.

Andrews, G., Stewart, G., Morris-Yates, A., Holt, P., & Henderson, S. (1990). Evidence for a general neurotic syndrome. *British Journal of Psychiatry*, *157*, 6–12.

Bailey, C. J., & Miller, N. E. (1952). The effect of sodium amytal on an approach–avoidance conflict in cats. *Journal of Comparative and Physiological Psychology*, *45*, 205–208.

Barry, H., & Miller, N. E. (1962). Effects of drugs on approach–avoidance conflict tested repeatedly by means of a telescope alley. *Journal of Comparative and Physiological Psychology*, *55*, 201–210.

Barry, H., & Miller, N. E. (1965). Comparison of drug effects on approach, avoidance and escape motivation. *Journal of Comparative and Physiological Psychology*, *59*, 18–24.

Carver, C. S. (2005). Impulse and constraint: Perspectives from personality psychology, convergence with theory in other areas, and potential for integration. *Personality and Social Psychology Review*, *9*, 312–333.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: Evidence and implications. *Psychological Bulletin*, *135*, 183–204.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*, 319–333.

Claridge, G. S. (1967). *Personality and arousal: A psychophysiological study of psychiatric disorder*. Oxford: Pergamon Press.

Cloninger, C. R. (1986). A unified biosocial theory of personality and its role in the development of anxiety states. *Psychiatric Developments*, *3*, 167–226.

Cloninger, C. R., Svrakic, D. M., & Przybecky, T. R. (1993). A psychobiological model of temperament and character. *Archives of General Psychiatry*, *50*, 975–990.

Corr, P. J. (2016a). *Hans Eysenck: A contradictory psychology*. London: Palgrave.

Corr, P. J. (2016b). Reinforcement sensitivity theory of personality questionnaires: Structural survey with recommendations. *Personality and Individual Differences*, *89*, 60–64.

Corr, P. J. (Ed.) (2008). *The reinforcement sensitivity theory of personality*. Cambridge: Cambridge University Press.

Corr, P. J., & Cooper, A. B. (2016). The Reinforcement Sensitivity Theory of Personality Questionnaire (RST-PQ): Development and validation. *Psychological Assessment*, *28*, 1427–1440.

Corr, P. J., & McNaughton, N. (2008). Reinforcement sensitivity theory and personality. In P. J. Corr (Ed.), *The reinforcement sensitivity theory of personality*. Cambridge: Cambridge University Press.

Corr, P. J., & McNaughton, N. (2012). Neuroscience and approach/avoidance personality traits: A two stage (valuation–motivation) approach. *Neuroscience and Biobehavioral Reviews*, *36*, 2339–2354.

Corr, P. J., & McNaughton, N. (2016). Neural mechanisms of low trait anxiety and risk for externalizing behaviour. In T. P. Beauchaine & S. P. Hinshaw (Eds.), *The Oxford handbook of externalizing spectrum disorders*. Oxford Handbooks Online: Oxford University Press.

Corr, P. J., Pickering, A., & Gray, J. A. (1995). Personality and reinforcement in associative and instrumental learning. *Personality and Individual Differences*, *19*, 47–71.

Davidson, R. J., Ekman, P., Saron, C. D., Senulis, J. A., & Friesen, W. V. (1990). Approach–withdrawal and cerebral asymmetry: Emotional expression and brain physiology I. *Journal of Personality and Social Psychology*, *58*, 330–341.

Davidson, R. J., Shackman, A. J., & Maxwell, J. S. (2004). Asymmetries in face and brain related to emotion. *Trends in Cognitive Sciences*, *8*, 389–391.

Dawkins, R. (2005). *The Ancestor's Tale: A pilgrimage to the dawn of life*. London: Phoenix (Orion Books Ltd).

Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation and extraversion. *Behavioral and Brain Sciences*, *22*, 491–569.

DeYoung, C. G. (2015). Cybernetic Big Five theory. *Journal of Research in Personality*, *56*, 33–58.

DeYoung, C. G., & Gray, J. R. (2009). Personality neuroscience: Explaining individual differences in affect, behaviour and cognition. In P. J. Corr & G. Matthews (Eds.), *The Cambridge handbook of personality psychology*. Cambridge: Cambridge University Press.

Elliot, A. J., & Thrash, T. M. (2002). Approach–avoidance motivation in personality: Approach and avoidance temperament and goals. *Journal of Personality and Social Psychology*, *82*, 804–818.

Elliot, A. J., & Thrash, T. M. (2010). Approach and avoidance temperament as basic dimensions of personality. *Journal of Personality*, *78*, 865–906.

Eysenck, H. J. (1963). Editorial. *Behaviour Research and Therapy*, *1*, 1–2.

Eysenck, H. J. (1967). *The biological basis of personality*. Springfield, IL: Charles C. Thomas.

Gardini, S., Cloninger, C. R., & Venneri, A. (2009). Individual differences in personality traits reflect structural variance in specific brain regions. *Brain Research Bulletin*, *79*, 265–270.

Gray, J. A. (1964a). Strength of the nervous system and levels of arousal: A reinterpretation. In J. A. Gray (Ed.), *Pavlov's typology*. Oxford: Pergamon Press.

Gray, J. A. (1964b). Strength of the nervous system as a dimension of personality in man. In J. A. Gray (Ed.), *Pavlov's typology*. Oxford: Pergamon Press.

Gray, J. A. (Ed.) (1964c). *Pavlov's typology*. Oxford: Pergamon Press.

Gray, J. A. (1967). Disappointment and drugs in the rat. *Advancement of Science*, *23*, 595–605.

Gray, J. A. (1970a). The psychophysiological basis of introversion–extraversion. *Behaviour Research and Therapy*, *8*, 249–266.

Gray, J. A. (1970b). Sodium amobarbital, the hippocampal theta rhythm and the partial reinforcement extinction effect. *Psychological Review*, *77*, 465–480.

Gray, J. A. (1971). *The psychology of fear and stress*. London: Weidenfeld and Nicolson.

Gray, J. A. (1972). Learning theory, the conceptual nervous system and personality. In V. D. Nebylitsyn & J. A. Gray (Eds.), *The biological bases of individual behaviour*. London, New York: Academic Press.

Gray, J. A. (1975). *Elements of a two-process theory of learning*. London: Academic Press.

Gray, J. A. (1976). The behavioural inhibition system: A possible substrate for anxiety. In M. P. Feldman & A. M. Broadhurst (Eds.), *Theoretical and experimental bases of behaviour modification*. London: Wiley.

Gray, J. A. (1977). Drug effects on fear and frustration: Possible limbic site of action of minor tranquilizers. In L. L. Iversen, S. D. Iversen, & S. H. Snyder (Eds.), *Handbook of psychopharmacology: Vol. 8 – Drugs, neurotransmitters and behaviour*. New York: Plenum Press.

Gray, J. A. (1982). *The neuropsychology of anxiety: An enquiry in to the functions of the septo-hippocampal system*. Oxford: Oxford University Press.

Gray, J. A. (1987). *The psychology of fear and stress*. London: Cambridge University Press.

Gray, J. A., & Ball, G. G. (1970). Frequency-specific relation between hippocampal theta rhythm, behavior and amobarbital action. *Science*, *168*, 1246–1248.

Gray, J. A., & McNaughton, N. (1983). Comparison between the behavioural effect of septal and hippocampal lesions: A review. *Neuroscience and Biobehavioral Reviews*, *7*, 119–188.

Gray, J. A., & McNaughton, N. (2000). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system* (2nd ed.). Oxford: Oxford University Press.

Gray, J. A., & Smith, P. T. (1969). An arousal-decision model for partial reinforcement and discrimination learning. In R. Gilbert & N. S. Sutherland (Eds.), *Animal discrimination learning*. London: Academic Press.

Hebb, D. O. (1955). Drives and the C.N.S. (conceptual nervous system). *Psychological Review*, *62*, 243–254.

Hengartner, M. P., Tyrer, P., Ajdacic-Gross, V., Angst, J., & Rossler, W. (2017). Articulation and testing of a personality-centred model of psychopathology: Evidence from a longitudinal community study over 30 years. *European Archives of Psychiatry and Clinical Neuroscience*, 1–12.

Hull, C. L. (1943). *Principles of behaviour*. New York: Appleton–Century–Crofts, Inc.

Hull, C. L. (1952). *A behavior system*. New Haven, CT: Yale University Press.

John, O. P., & Srivastava, S. (1995). The Big Five trait taxonomy: History, measurement and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: theory and research* (2nd ed.). London: Guilford Press.

Kimble, G. A. (1961). *Hilgard and Marquis' conditioning and learning*. New York: Appleton–Century–Crofts.

McNaughton, N. (2017). What do you mean 'anxiety'? Developing the first anxiety syndrome biomarker. *Journal of the Royal Society of New Zealand*, *48*, 177–190.

McNaughton, N., & Corr, P. J. (2004). A two-dimensional neuropsychology of defense: Fear/anxiety and defensive distance. *Neuroscience and Biobehavioral Reviews*, *28*, 285–305.

McNaughton, N., & Corr, P. J. (2014). Approach, avoidance, and their conflict: The problem of anchoring. *Frontiers in Systems Neuroscience*, *8*.

McNaughton, N., & Corr, P. J. (2016). Mechanisms of comorbidity, continuity, and discontinuity in anxiety-related disorders. *Development and Psychopathology*, *28*, 1053–1069.

McNaughton, N., DeYoung, C. G., & Corr, P. J. (2016). Approach/avoidance. In J. R. Absher & J. Cloutier (Eds.), *Neuroimaging personality, social cognition and character*. San Diego, CA: Elsevier.

Miller, N. E. (1944). Experimental studies of conflict. In J. M. Hunt (Ed.), *Personality and the behavioural disorders*. New York: Ronald Press.

Miller, N. E. (1959). Liberalization of basic S-R concepts: Extensions to conflict behaviour, motivation and social learning. In S. Koch (Ed.), *Psychology: A study of a science*. New York: Wiley.

Mowrer, O. H. (1960). *Learning theory and behavior*. New York: Wiley.

Pavlov, I. P. (1927). *Conditioned reflexes* (trans. G. V. Anrep). London: Constable & Co. Ltd, Dover edition, by special arrangement with Oxford University Press.

Pavlov, I. P. (1932). The reply of a physiologist to psychologists. *Psychological Review*, *39*, 91–127. (Reprinted in Kaplan, M. (1966) *Essential works of Pavlov*. London: Bantam Books.)

Pickering, A. D., Corr, P. J., & Gray, J. A. (1999). Interactions and reinforcement sensitivity theory: A theoretical analysis of Rusting & Larsen (1997). *Personality and Individual Differences*, *26*, 357–365.

Rapoport, J. L. (1989). The biology of obsessions and compulsions. *Scientific American*, *260* (March), 63–69.

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

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

Wilson, G. D., Gray, J. A., & Barrett, P. T. (1990). A factor analysis of the Gray–Wilson personality questionnaire. *Personality and Individual Differences*, *11*, 1037–1045.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*, 459–482.

Zald, D., & Depue, R. (2001). Serotonergic modulation of positive and negative affect in psychiatrically healthy males. *Personality and Individual Differences*, *30*, 71–86.