Abstract

Cognitive neuroscience has developed many approaches to the study of learning that might be useful to functionally oriented researchers, including those from a relational frame theory (RFT) perspective. We focus here on two examples. First, cognitive neuroscience often distinguishes between habit and goal-directed reinforcement learning, in which only the latter is sensitive to proximal changes in behavior-environment contingencies. This distinction is relevant to RFT's original concerns about how rule-based processes can sometimes render an individual's behavior maladaptive to changing circumstances. Second, the discovery of neurophysiological structures associated with fear extinction and generalization can potentially yield new insights for derived relational responding research. In particular, we review how such work not only informs new ways of modifying the functions transformed in derived relational responding, but also new ways of measuring derived relational responding itself. Overall, therefore, existing conceptual and methodological advances in the cognitive neuroscience literature addressing learning appear to generate functionally interesting predictions related to RFT that might not have surfaced from a traditional functional analysis of behavior.

The goal of cognitive neuroscience is to understand how cognitive activities emerge from biological operations in neural tissue. At first, this field might seem far removed from functional psychology where behavioral events are explained relative to measurable regularities within the environment (De Houwer, Barnes-Holmes & Moors, 2013). Yet cognitive neuroscience can meaningfully contribute to the progression of functional psychological endeavors, including relational frame theory (RFT; Hayes, Barnes-Holmes & Roche, 2001; Vahey & Whelan, 2016). This is possible because cognitive neuroscience measures behavioral events as a proxy for cognitive activity (e.g. De Houwer, 2011; De Houwer, Gawronski, & Barnes-Holmes, 2013, p. 16). These behavioral events include brain activity, typically assayed non-invasively in humans using functional magnetic resonance imaging (fMRI) or electroencephalography (EEG). Here, we focus on two recent conceptual and methodological developments in the cognitive neuroscience literature. The first posits that it is possible to measure the relative contribution to learning of, broadly speaking, verbal behavior versus direct behavior-environment contingencies. The second development is concerned with both the extinction and generalization of conditioned fear. These advances in the cognitive neuroscience of learning generate interesting functional predictions related to RFT that might not have surfaced from a traditional functional analysis of behavior.

Goal-directed vs. habitual behavior

A core concept in RFT is that behavior does not need to be under the control of directly present behavior-environment contingencies. Indeed, verbal behavior (i.e., controlled by contingencies not directly present) can sometimes be at odds with the history of reinforcement experienced by an individual. For example, if a person with arachnophobia is told that there is the possibility of a spider inside a packet of their favorite snack, that person is highly unlikely to reach into the packet despite a long history of reinforcement for that behavior. In a broadly analogous manner, cognitive neuroscience often distinguishes goal-directed behavior from habitual behavior. Goal-directed behavior here refers to the ability to strategically calculate and then select the *optimal* action for obtaining a given outcome based on the current outcome value assigned to each response alternative (as a result of previous learning trials). More specifically, goal-directed behavior is posited to work using a learned internal model of the world (e.g., by representing, testing and updating possible action outcomes) with values encoded in the ventromedial prefrontal cortex (vmPFC; part of the executive system) via the dorsomedial striatum (namely, the caudate nucleus). In contrast, habitual behavior is viewed as being gradually 'stamped in' to the vmPFC via the dorsolateral striatum (namely, the putamen) by the overall history of reinforcement for a particular response so that it tends to be emitted regardless of proximal changes in the behavior-environment contingencies that originally influenced it (see Doll, Simon, & Daw, 2012; Wood & Rünger, 2016, pp. 291-292, 298-301). Thus, to describe behavior as being habitual is to describe it as being in some sense automatic *and* inflexible (i.e., *cognitively efficient*; see De Houwer & Moors, 2012; Wood & Rünger 2016). The

extant cognitive neuroscience literature suggests that both types of learning operate in parallel, and the degree to which a behavior is under the relative control of goal-directed versus habitual processes is influenced by the degree to which that behavior has been repeatedly and consistently reinforced in the relevant context (Daw, Gershman, Seymour, Dayan & Dolan, 2011; O'Hare et al., 2016; Redgrave et al., 2010). The relative influence of model-based over habitual learning has also been shown to develop with age (Decker, Otto, Daw, Hartley, 2016).

The RFT literature describes experimental tasks that require participants to derive new stimulus relations in the absence of competing reinforcement contingencies. In deriving these relations, verbally sophisticated participants likely engage in trial and error rule-generation during the training phases of such tasks until they reliably obtain reinforcement for responding in accordance with a given self-generated response rule (see Cabello, Luciano, Gomez & Barnes-Holmes, 2004; Hayes, White & Bissett, 1998; Luque & O'Hora, 2016). As such, the empirical studies underpinning RFT have generally focused on the emergence of rule-based behavior that is optimally sensitive to environmental regularities (i.e., analogous to goal-directed processes in the cognitive neuroscience literature), without necessarily accounting for the influence of relatively direct contingency learning. However, it has recently been argued that human behavior should not be conceptualized as being either exclusively verbal or exclusively non-verbal but rather as being on a continuum between these two extremes (Barnes-Holmes, Barnes-Holmes, Hussey & Luciano, 2016; Hughes, Barnes-Holmes & Vahey, 2012, pp. 32-34). The *Multi-dimensional Multi-level* (MDML) analytic framework identifies four distinct ways in which a given behavior might lie on the continuum between verbal and non-verbal: behavior can be more versus less *derived*, *relationally complex*, *relationally coherent* or most importantly for the present purposes – *relationally flexible* (see Barnes-Holmes et al., 2016, pp. 123-124). Barnes-Holmes et al. defined relational flexibility as a property that describes behavior to be "more or less sensitive to current contextual variables" (p. 123) – acknowledging that this variable awaits systematic description via experimental research.

An exciting avenue for future research could be to investigate the conditions under which rule-based behavior is more or less insensitive to reinforcement contingencies (O'Hora, Barnes-Holmes, & Stewart, 2014). Against this backdrop, the already well-established literature underpinning the cognitive neuroscience of habits is likely to be a useful source of both data and ideas for the theoretical development of RFT. Regarding data, the habit literature contains a large store of functional findings that are compatible with, but generally not considered by, RFT, including a large body of research findings describing what functional features lead operant behavior to become relatively habitual over time (see Wood & Rünger, 2016). For example, instrumental behavior is particularly likely to become habitual when it is relatively (a) uncomplicated, and (b) consistently reinforced in a (c) stable context such that reinforcement is delivered (d) intensively and (e) for a relatively prolonged period of time (Adams & Dickinson, 1981; Ostlund & Balleine, 2009; Wood &

Rünger, 2016, p. 295). In addition, instrumental behavior is particularly likely to become habitual under (e) variable interval schedules of reinforcement, and/or when the relevant individual is (f) stressed, (g) under the influence of psychostimulants (e.g., tobacco or cocaine), and/or (h) disinclined to deliberate about the relevant behavior (Gillan, Otto, Phelps, & Daw, 2015; Wood & Rünger 2016, pp. 295-305).

There is even preliminary evidence that (i) avoidance behavior is more prone to becoming habitual than behavior that is positively reinforced (likely because it is more difficult, by definition, for an individual to discriminate when an avoidance contingency has been discontinued or otherwise devalued; see Holland, 2008, pp. 239-40; LeDoux, Moscarello, Sears, & Campese, 2017, pp. 26-31; Gillan, Urcelay, & Robbins, 2016). This ambition to develop measures of the individual's tendency toward habitual and non-deliberative avoidance (e.g. Gillan et al., 2011; Gillan et al., 2014) is clearly complementary with RFT's traditional focus upon *experiential/psychological avoidance* as a deliberative form of avoidance implicated in a diverse array of psychopatholoical conditions (e.g., Hayes, Wilson, Gifford, Follette, & Strosahl, 1996). For example, Gillan et al. (2014) found that unlike controls, participants diagnosed with obsessive-compulsive disorder persisted with a modestly trained pedal-press habit for avoiding an electric shock to the wrist even after they were extensively instructed, shown and reported believing that the electrode attached to his/her wrist had been disconnected from the relevant electric power source. Crucially, even though all of Gillan et al.'s (2014) participants with OCD clearly derived the intended meaning of those instructions, this often failed to disrupt the relevant pedal-press avoidance function (i.e., unlike the control participants for whom the relevant function was completely disrupted by the instructions). Such examples raise key questions for RFT about when derived relational responding is likely to be successful versus not in disrupting or otherwise transforming the functions of habitual behavior (e.g., whether by transforming the functions of its cues or its consequences). By systematically examining those situations in which some people display compulsive (or addictive) behaviors RFT may discover useful answers to such questions.

In summary, cognitive neuroscience models of reinforcement learning clearly have potential to inspire meaningful conceptual development of RFT with respect to the factors that render derived relational responding more or less influential in moderating directly contacted operant or respondent processes (and vice versa). Moreover, as per Yin & Knowlton (2006, p. 474), there is clearly a longstanding appetite among neuroscientists to elaborate their models of habitual behavior based upon the kinds of functional distinctions made by RFT:

> *Given the enormous structural complexity of the basal ganglia, a strictly bottom-up approach in elucidating their [habit-related] functions might not be fruitful. Instead, research can be guided by a top-down analysis based upon behavior.*

Thus, such research has the potential to not only yield a wider audience for RFT, but also to yield substantial applied value insofar as it seeks to provide a technical understanding of what environmental variables govern our ability to modify relatively entrenched and problematic behavioral patterns deliberately via derived relational responding.

Fear learning: A view from cognitive neuroscience, and directions for RFT and behavior analysis

Pavlovian conditioning is widely used in the study of human fear learning (Beckers, Krypotos, Boddez, Effting, & Kindt, 2013). First, a neutral stimulus (e.g. a tone) is repeatedly paired with an unpleasant outcome (unconditioned stimulus; US, e.g. a brief electric shock). The antecedent stimulus is then likely to evoke preparatory defense responses, both physiological (e.g. enhanced skin conductance or eye-blink startle potentiation) and behavioral (e.g. conditioned suppression); at this stage, the once neutral stimulus is referred to as a conditioned stimulus (CS). Operant conditioning is often incorporated into these paradigms when US omission is made contingent on the production of an overt response (e.g. a specific arm movement or rate of key-pressing) (Lovibond, 2006; Meulders, Franssen, Fontyne & Vlaeyen, 2016; Vervliet & Indekeu, 2015). In general, these learning processes might reflect emergent fear and avoidance responding following a real-world aversive experience (Craske, Hermans & Vansteenwegen, 2006).

Cognitive neuroscience typically views fear-learning processes from the 'implementational' and 'algorithmic' perspectives (Marr, 1982; Moors, 2007). The former approach describes processes with respect to necessary and underlying neurophysiological mechanisms (e.g. Davis, 1992), while the latter approach focuses on the meditational role of representational structures (e.g. Bouton, 2007). It is assumed, for example, that sensory information about a CS and US are independently stored in memory. These traces are both activated during contingent CS-US pairings, which allows an associative link to form between the two (Hall, 1996). This CS-US association converges in the basolateral amygdala (BLA) and is excited by projections from the sensory cortices in response to any subsequent CS presentation (LaBar, Gatenby, Gore, LeDoux & Phelps, 1998; LeDoux, 1998). The BLA then engages the central nucleus of the amygdala (Ce), which is the primary output nucleus for conditioned fear responses (Wilensky, Schafe, Kristensen & LeDoux, 2006). Projections from Ce to hypothalamus are related with increased autonomic arousal, and the prelimbic (PL) prefrontal cortex and ventral striatum (VS) are implicated in the emission of passive and active avoidance behavior (Bravo-Rivera, Roman-Ortiz, Brignoni-Perez, Sotres-Bayon, & Quirk, 2014; Davis, 1992; Lissek, 2012; Sotres-Bayon & Quirk, 2010). Cognitive neuroscience therefore generates information about the internal neurophysiology associated with fear expression, as well as a framework to describe the mental structures that translate into fear-related behavioral output (e.g. Bravo-Rivera, et al. 2015; Büchel, Morris, Dolan & Friston, 1998; Lissek et al., 2014; Moors, 2009).

In contrast to cognitive neuroscience, RFT and behavior analysis use functional definitions (e.g., Dymond & Roche, 2009; Hayes et al., 1996), in which learning processes are explained with respect to a history of measurable correlations between environmental stimuli and behavioral responses (Skinner, 1974), and this mode of analysis produces information about the circumstances associated with fear (e.g. Dinsmoor, 2001; Friman, Hayes & Wilson, 1988; Lohr, Olatungi, & Sawchuk, 2007; Mowrer, 1939; Weiner, 1963). But a judicious exploration of the concepts and methods from cognitive neuroscience may help to further the depth of clinical RFT and behavior analysis (e.g. De Houwer, 2011; Hayes, Barnes-Holmes & Wilson, 2012; Zentall, 2012). We first indicate how cognitive neuroscience has revealed new functional information about the extinction of conditioned fear, which is a psychological phenomenon that is frequently discussed in the RFT literature (e.g. Blackledge, 2007; Dougher, Augustson, Markham, Greenway, & Wulfert, 1994; Roche, Kanter, Brown, Dymond & Fogarty, 2008). Similarly, we propose that cognitive neuroscience offers insights into the generalization of fear, another psychological phenomenon discussed in the RFT literature (e.g. Auguston & Dougher, 1997).

Cognitive neuroscience and extinction learning. A functional definition of extinction is 'a decreased frequency of a conditioned response (CR; e.g. heightened autonomic arousal or 'freezing') by virtue of the repeated presentation of a CS (e.g. a tone) without the US (e.g. a brief electric shock)'. As such, extinction learning is described as a psychological phenomenon where conditioned fear simply diminishes with each CS presentation. A critical observation, however, is that extinction does not totally erase conditioned fear from an organism's repertoire. A return of extinguished fear is often seen following (i) the passage of time (spontaneous recovery; Pavlov, 1927), (ii) a change in context (contextual renewal; Bouton & King, 1983) and (iii) a re-presentation of the US (reinstatement; Hermans et al., 2005). The basic functional definition of extinction does not readily tell us about the factors relating to robust extinction learning. That is, the basic definition does not contain information as to the environmental conditions associated with the return-of-fear after initial extinction. Supplementing a functional account with recent evidence from cognitive neuroscience (e.g., Onat & Büchel, 2015) can both provide added information about the functional conditions that promote extinction and afford a more complete theory of extinction that accounts for post-extinction return of fear effects (e.g. Milad & Quirk, 2012).

It is often assumed within cognitive neuroscience literature that an original, 'excitatory' association between the CS-US remains intact in memory while a secondary 'inhibitory' association is formed during extinction; this inhibitory association encodes that the CS no longer predicts the US. These dichotomous associations compete with one another, thus establishing a CS whose function is ambiguous and determined by additional environmental factors. Recall of the inhibitory CS-US association might fail, for example, if the CS appears in a context that is dissimilar to the one where extinction took place, leading to return-of-fear (Bouton, 2002; Bouton & King, 1983; Bouton,

Winterbauwer & Todd, 2012; Pavlov, 1927). This is known as the inhibitory learning model of extinction; it is an account that reframes extinction as an active learning process. Infralimbic (IL) and ventromedial prefrontal cortex (vmPFC) activity has, in particular, been implicated in the retention of extinction after a delayed period. Stated otherwise, inhibitory associations converge on the IL for long-term storage and, once excited, the IL dampens conditioned fear via projections to amygdala (Milad & Quirk, 2012; Quirk, Garcia, Gonzalez-Lima, 2006). Indeed, IL neurons have been observed to selectively activate in response to an extinguished CS in the days after extinction and the level of activity negatively correlates with the expression of fear (Milad & Quirk, 2002). Lesions to the vmPFC that center on IL are also associated with a return of fear in the days after successful extinction (Quirk, Russo, Barren, & Lebron, 2000). Furthermore, evidence suggests that extinction-related vmPFC activity is modulated by input from the hippocampus, which is associated with the retention of context-specific sensory information (Corcoran & Quirk, 2007). Cognitive neuroscience has therefore broken new ground and uncovered internal mechanisms that may account for extinction learning and return-of-fear. In short, the potential for the extinguished CS to evoke fear is retained by the structures surrounding the BLA and Ce, but this circuitry can be inhibited by contextually dependent activity within the IL, vmPFC and hippocampus.

Experimental research from the cognitive neuroscience perspective has the potential to augment our original functional understanding of extinction, thus paving the way to optimized extinction learning and restricted return-of-fear (see Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014). Haddad and colleagues (2015), for example, recently hypothesized that differences in PFC maturation during adolescence might impact on the ability to form inhibitory CS-US associations. Their investigation revealed that anxious adolescents indeed elicited heightened conditioned fear to a safety cue that was explicitly paired with an omitted US (i.e. a CS-), and brain activity in areas such as the dlPFC were negatively correlated with age. In contrast, non-anxious adolescents did not elicit heightened responding to the CS, and brain activity in areas such as the vmPFC were instead positively correlated with age (Haddad, Bilderbeck, James, & Lau, 2015). This evidence suggests that inhibitory-based extinction learning may improve with age but the developmental trajectory is perturbed in those with high levels of anxiety. In addition, and building upon the inhibitory-learning model of extinction, Vansteenwegen and colleagues (2007) highlighted a means to attenuate return-of-fear. It was predicted that installing new inhibitory association across multiple contexts could interfere with the re-activation of the original CS-US associations in a novel context (also, see Neumann, 2006). Phobic students were, therefore, repeatedly shown videos of a spider in either one specific location or three different locations. Fear responding to the spider reliably reduced across extinction trials and a return-of-fear was observed when the spider was then presented in a new context. However, the return-of-fear was significantly diminished in the multiple locations extinction-

group (Vansteenwegen et al., 2007). This finding suggests that that extinction learning can be functionally enhanced when the CS is extinguished across multiple contexts.

Advances in cognitive neuroscience clearly highlight new information about the conditions related to extinction learning, which have implications for on-going clinical RFT research. Vervoort and colleagues (2014), for example, recently demonstrated restrictions in the transformation of extinction functions through stimulus equivalence classes. On the one hand, extinguishing fear responding to a CS+ resulted in derived fear reduction to other members of a stimulus equivalence class. On the other hand, extinguishing fear responding to an equivalent stimulus bared no discernable impact on fear responding to the CS+ (Vervoort, Vervliet, Bennett & Baeyens, 2014). This evidence suggests that the extinguished stimulus is a critically important factor in derived extinction of conditioned fear. Future research might now examine whether individual differences in age and anxiety impact on derived extinction (as indicated by Haddad, et al., 2015). It might be the case, for example, that derived extinction is diminished in adolescents and hampered even further in those with a prior history of anxious symptomology. This question is clinically interesting given that derived extinction learning is one (of many) mechanism-of-change that supposedly drives Acceptance and Commitment Therapy (ACT; see Blackledge, 2007; Dymond & Roche, 2009). If indeed derived extinction is weakened in anxious adolescents, then this population might be at risk for poorer therapeutic outcome and could benefit from tailored forms of ACT and behavior therapy (e.g. Craske et al., 2014). In addition, it may be interesting to examine whether the steps taken to maximize extinction learning also translate into derived extinction learning. Recent research has confirmed, for example, that the return of derived avoidance in a novel context can be greatly reduced by presenting the CS without the US across multiple contexts (Bennett, Roche, Baeyens, Broothaerts & Hermans, 2016).

Cognitive neuroscience and category-based fear learning

Humans do not just respond to the sensory properties of a fear-relevant event: conceptual knowledge of the interrelations between stimuli is also drawn upon during an aversive learning experience (Dunsmoor & Murphy, 2015). This is particularly obvious when abstract events evoke excessive fear and avoidance despite never featuring in a conditioning episode (Bennett, Vervoort, Boddez, Baeyens & Hermans, 2015). For instance, an individual who survives a traumatic car accident might later experience distress when confronted with symbols of driving (e.g., the sound of keys) and even avoid other modes of transport (e.g., trains or boats; Yule, Bolton, Udwin, Boyle, O'Ryan, & Nurrishh, 2000). RFT research has experimentally replicated the learning history that might lead to this sort of outcome. An artificial verbal category is typically established by way of a stimulus equivalence class (e.g. A1=B1=C1=D1) and a US is repeatedly paired with a single member (i.e., B1+). The behavioral control exercised by other members of this stimulus equivalence class (i.e.,

C1 & D1) subsequently alters (or transforms) as they evoke heightened fear-related responses; this is known as a *transformation of stimulus function* (Hayes, et al., 2001).

There is increased recognition of the role that conceptual knowledge plays in fear acquisition within the cognitive neuroscience literature and, here, the term 'category-based generalization' is common (Dunsmoor & Murphy, 2015). The literature documents altered behavioral responding to stimuli that are categorically related to an aversively conditioned stimulus, and changes in the neurophysiological structures that hypothetically encode for category-selective information. For example, Dunsmoor and colleagues posited that the neuronal representation of an entire category is modified when single exemplars are associated with an aversive US and this neurophysiological change may facilitate a transfer of fear to previously neutral exemplars (Dunsmoor, Kregal, Martin & LaBar, 2014). Therefore, changes within and around category-selective regions of the occipitotemporal cortex were examined during fear conditioning. In this paradigm, forty different exemplars from one category (e.g., 'types of animals') were paired with a brief electric shock (US) and 40 different exemplars from the other category (i.e., 'types of tools') were paired with an omitted US. Participants generally elicited heightened skin conductance and self-reported US expectancy to exemplars of the aversively conditioned category, relative to exemplars of the non-aversively conditioned category. The affective meaning of an entire category clearly shifted as novel exemplars evoked conditioned fear. A neurophysiological analysis concurrently targeted category-selective regions of interest (ROI); namely, the lateral fusiform gyrus, posterior superior temporal sulcus and inferior occipital regions, which activate in response to images of animals (or animate stimuli), and the medial fusiform gyrus and posterior middle temporal gyrus, which activate in response to images of tools (or inanimate stimuli) (see Dunsmoor, et al. 2014). Aversive conditioning was associated with increased functional activity in the category-selective brain regions as a function of the Pavlovian contingency: animal-selective brain regions enhanced when individual animals were associated with the US while tool-selective brain regions enhanced when individual tools were associated with the US. Activity patterns within a set of voxels from the occipitotemporal cortex were also compared amongst the different category exemplars; this is known as 'representational similarity analysis'. Voxel activation (or representational similarity) was most similar between exemplars from the aversively conditioned category. That is, activity patterns were most similar between animal-animal stimulus pairs (relative to animal-tool pairs or tool-tool pairs) when animals were paired with the US and activity patterns were most similar between tool-tool stimulus pairs (relative to animal-tool pairs and animal-animal pairs) when tools were paired with the US. Dunsmoor and colleagues' (2014) paradigm can be seen as similar to the fear learning research that features in the RFT literature (e.g., Augustson & Dougher, 1997; Dymond et al., 2011); a number of studies now show that participants elicited spontaneous fear to previously neutral stimuli by virtue of their arbitrary relation to a set of

physically distinct, directly conditioned stimuli. This is known as *the transformation of stimulus function*.

Neurophysiological evidence can provide additional *in vivo* insights into verbal learning processes. First, activity in category-selective brain regions could provide a novel way to measure the transformation of stimulus function. For instance, a 'representational similarity analysis' revealed that pairs of physically dissimilar stimuli elicited similar occipitotemporal cortex activity when they were conceptually similar and shared a functional outcome. Relative activity in category-selective ROIs could therefore be used to measure the emergent, functional interchangeability between physically dissimilar stimuli; this is of course a definitive component of stimulus equivalence. Behaviorally, the interchangeability between verbal stimuli is checked in two ways. Unreinforced conditional discriminations can be examined during an arbitrary testing phase of a Matching-to-Sample task and this is referred to as a 'derived relational response' (e.g. if A1→B1 & if A1→C1, then B1→C1 & C1→B1). Spontaneous Pavlovian or instrumental responding can be examined during a conditioning paradigm and this is specifically referred to as a 'transformation of stimulus function' (e.g. if B1 controls avoidance, then C1 will also control avoidance). We now suggest that neurophysiological activity in category-selective ROIs may offer a new means to index the functional interchangeability of verbal stimuli

Neurophysiological changes in category-selective ROIs could also act as a dynamic test for the transformation of stimulus function that provides information about speed and effort. Dunsmoor and colleagues, for example, demonstrated that previously novel category-exemplars elicited heightened fear and increased category-selective ROI activity once a number of other exemplars were paired with a US (Dunsmoor et al., 2014). This finding could suggest that the evaluative function of an entire verbal category transforms during a limited number of CS-US pairings. That is, the transformation of stimulus function could be already evident on a neurophysiological level in the seconds after a fear-conditioning episode. This approach contrasts with the *status quo* wherein the transformation of stimulus function is only said to occur after a member of a stimulus equivalence class (e.g. C1) evokes an overt behavioral response that was directly conditioned to another member (e.g. B1) (see Dymond & Roche, 2009). Functional imaging could therefore enhance the temporal sensitivity of RFT research; indeed, there are published examples of this approach (e.g., Amd, Barnes-Holmes & Ivanoff, 2013). Regarding the strength of the transformation of stimulus function, future research could examine individual differences in the activity of category-selective brain regions. For instance, do individuals with specific phobias show differences in brain activity in category-specific brain regions relative to healthy controls? Such temporal and spectral information will, in principle, drive novel research questions as well as create new clinical applications for RFT and behavioral analysis. As an example, Whelan and colleagues (2012) found that individual differences in

neurophysiological activity could differentiate among a range of phenotypes (substance misuse, ADHD symptoms), whereas overt behavioral measurements such as reaction time could not.

Conclusion

In the current paper we have reviewed various examples of how recent conceptual and empirical developments in cognitive neuroscience can foster the conceptual and empirical development of RFT in ways not previously anticipated from within the RFT literature. Initially, we focused upon how the cognitive neuroscience of habits highlights the need within RFT to consider how derived relational responding renders behavior more or less sensitive to behavior-environment contingencies. Then, we reviewed how the cognitive neuroscience of fear extinction and generalization could highlight new ways of measuring and modifying the functions transformed by derived relational responding. We therefore recommend the cognitive neuroscience literature to RFT researchers as a potential source of research innovation that extends far beyond the examples in this article. Of course, the uncritical adoption of cognitive neuroscience concepts is not recommended. However, whenever cognitive neuroscience overlaps with RFT in terms of its functional domain, RFT researchers have the opportunity to obtain ready-made information about the physiological constraints that apply to any given behavioral function (Vahey & Whelan, 2016). Indeed, as we have also argued throughout the current manuscript, it is possible that measures such as fMRI and EEG may be more sensitive than measures such as reaction time for detecting and thus discovering subtle functional relations between behavior and its environment, and may therefore add not only to the analytic depth of RFT but also to its precision and scope.

Acknowledgements

Dr Bennett was funded by an Irish Research Council postdoctoral fellowship grant (GOIPD/2016/617).

References

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, *33*(2), 109-121.

Amd, M., Barnes-Holmes, D., & Ivanoff, J. (2013). A derived transfer of eliciting emotional functions using differences among electroencephalograms as a dependent measure. *Journal of the Experimental Analysis of Behavior, 99(3)*, 318-334.

Augustson, E. M., & Dougher, M. J. (1997). The transfer of avoidance evoking functions through stimulus equivalence classes. *Journal of Behavior Therapy and Experimental Psychiatry, 28*, 181-191.

Barnes-Holmes, D., Barnes-Holmes, Y., Hussey, I., & Luciano, C. (2016). Relational frame theory: Finding its historical and intellectual roots and reflecting upon its future development. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *Handbook of contextual behavioral science* (pp. 117-128). New York, NY: Wiley-Blackwell.

Beckers, T., Krypotos, A. M., Boddez, Y., Effting, M., & Kindt, M. (2013). What's wrong with fear conditioning? *Biological Psychology, 92*, 90-96.

Bennett, M., Vervoort, E., Boddez, Y., Hermans, D., & Baeyens, F. (2015). Perceptual and conceptual similarities facilitate the generalization of instructed fear. *Journal of Behavior Therapy and Experimental Psychiatry, 48,* 149-155.

Bennett, M.P., Roche, B., Baeyens, F., Broothaerts, K., B., & Hermans, D. (2016). A Process-level Analysis of Cognitive Defusion. *Manuscript in preparation*.

Blackledge, J. T. (2007). Disrupting verbal processes: Cognitive defusion in acceptance and commitment therapy and other mindfulness-based psychotherapies. *The Psychological Record, 57*, 555-577.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. *Biological Psychiatry, 52*, 976-986.

Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis*. Sinauer Associates.

Bouton, M. E., & King, D. A. (1983). Contextual control of the extinction of conditioned fear: tests for the associative value of the context. *Journal of Experimental Psychology: Animal Behavior Processes, 9*, 248-265.

Bouton, M. E., Winterbauer, N. E., & Todd, T. P. (2012). Relapse processes after the extinction of instrumental learning: renewal, resurgence, and reacquisition. *Behavioural Processes, 90(1)*, 130-141.

Bravo-Rivera, C., Roman-Ortiz, C., Brignoni-Perez, E., Sotres-Bayon, F., & Quirk, G. J. (2014). Neural structures mediating expression and extinction of platform-mediated avoidance. *The Journal of Neuroscience, 34*, 9736-9742.

Büchel, C., Morris, J., Dolan, R. J., & Friston, K. J. (1998). Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron, 20*, 947-957.

Cabello, F., Luciano, C., Gomez, I., & Barnes-Holmes, D. (2004). Human schedule performance, protocol analysis, and the" silent dog" methodology. *The Psychological Record*, *54*(3), 405-422.

Corcoran, K. A., & Quirk, G. J. (2007). Activity in prelimbic cortex is necessary for the expression of learned, but not innate, fears. *The Journal of Neuroscience, 27*, 840-844.

Craske, M. G., Hermans, D. E., & Vansteenwegen, D. E. (2006). *Fear and learning: From basic processes to clinical implications.* American Psychological Association.

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10-23.

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual Review of Neuroscience, 15*, 353-375.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848-858.

De Houwer, J. (2011). Why the cognitive approach in psychology would profit from a functional approach and vice versa. *Perspectives on Psychological Science, 6*, 202-209.

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review, 20*, 631-642.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology*, *24*(1), 252-287.

De Houwer, J., & Moors, A. (2012). How to define and examine implicit processes? In R. Proctor & E. J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes* (pp. 183-198). Oxford: Oxford University Press.

Dinsmoor, J. A. (2001). Stimuli inevitably generated by behavior that avoids electric shock are inherently reinforcing. *Journal of the Experimental Analysis of Behavior, 75,* 311-333.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*, 1075-1081.

Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior, 62*, 331-351.

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: how humans generalize fear. *Trends in Cognitive Sciences, 19*, 73-77.

Dunsmoor, J. E., Kragel, P. A., Martin, A., & LaBar, K. S. (2014). Aversive learning modulates cortical representations of object categories. *Cerebral Cortex, 24*, 2859-2872.

Dymond, S., & Roche, B. (2009). A contemporary behavioral analysis of anxiety and avoidance. *The Behavior Analyst, 32*, 7-28.

Dymond, S., Schlund, M. W., Roche, B., Whelan, R., Richards, J., & Davies, C. (2011). Inferred threat and safety: Symbolic generalization of human avoidance learning. *Behaviour Research and Therapy, 49*, 614-621.

Friman, P. C., Hayes, S. C., & Wilson, K. G. (1998). Why behavior analysts should study emotion: The example of anxiety. *Journal of Applied Behavior Analysis, 31*, 137-156.

Gillan, C. M., Morein-Zamir, S., Urcelay, G. P., Sule, A., Voon, V., Apergis-Schoute, A. M.,…Robbins, T. W. (2014). Enhanced avoidance habits in obsessive-compulsive disorder. *Biological psychiatry*, *75*(8), 631-638.

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523-536.

Gillan, C. M., Papmeyer, M., Morein-Zamir, S., Sahakian, B. J., Fineberg, N. A., Robbins, T. W., & de Wit, S. (2011). Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *American Journal of Psychiatry*, *168*(7), 718-726.

Gillan, C. M., Urcelay, G. U. & Robbins, T. W. (2016). An associative account of avoidance. In R. A. Murphy & R. C. Honey (Eds.), *Handbook of cognitive neuroscience and learning* (pp. 442-467). Wiley and Sons.

Haddad, A. D., Bilderbeck, A., James, A. C., & Lau, J. Y. (2015). Fear responses to safety cues in anxious adolescents: Preliminary evidence for atypical age-associated trajectories of functional neural circuits. *Journal of Psychiatric Research, 68*, 301-308.

Hall, G. (1996). Learning about associatively activated stimulus representations: Implications for acquired equivalence and perceptual learning. *Animal Learning & Behavior, 24,* 233-255.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Springer Science & Business Media.

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual behavioral science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science*, *1*(1), 1-16.

Hayes, S. C., White, D., & Bissett, R. T. (1998). Protocol analysis and the "silent dog" method of analyzing the impact of self-generated rules. *The Analysis of Verbal Behavior*, *15*, 57-63.

Hayes, S. C., Wilson, K. G., Gifford, E. V., Follette, V. M., & Strosahl, K. (1996). Experiential avoidance and behavioral disorders: A functional dimensional approach to diagnosis and treatment. *Journal of Consulting and Clinical Psychology, 64*, 1152-1168.

Hermans, D., Dirikx, T., Vansteenwegenin, D., Baeyens, F., Van den Bergh, O., & Eelen, P. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy, 43*, 533-551.

Holland, P. C. (2008). Cognitive versus stimulus-response theories of learning. *Learning & Behavior*, *36*(3), 227-241.

Hughes, S., Barnes-Holmes, D., & Vahey, N. (2012). Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. *Journal of Contextual Behavioral Science*, 1(1), 17-38.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex, 13*, 400-408.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron, 20*, 937-945.

LeDoux, J. (1998). Fear and the brain: where have we been, and where are we going? *Biological Psychiatry, 44*, 1229-1238.

LeDoux, J., Moscarello, J., Sears, R., & Campese, V. (2017). The birth, death and resurrection of avoidance: A reconceptualization of a troubled paradigm. *Molecular Psychiatry*, *22*(1), 24-36.

Lissek, S. (2012). Toward an account of clinical anxiety predicated on basic, neurally mapped mechanisms of Pavlovian fear-learning: The case for conditioned overgeneralization. *Depression and Anxiety, 29*, 257-263.

Lissek, S., Bradford, D. E., Alvarez, R. P., Burton, P., Espensen-Sturges, T., Reynolds, R. C., & Grillon, C. (2014). Neural substrates of classically conditioned fear-generalization in humans: a parametric fMRI study. *Social Cognitive and Affective Neuroscience, 9,* 1134-1142.

Lohr, J. M., Olatunji, B. O., & Sawchuk, C. N. (2007). A functional analysis of danger and safety signals in anxiety disorders. *Clinical psychology review, 27*, 114-126.

Lovibond, P. (2006). Fear and avoidance: An integrated expectancy model. In M. Craske, D. Hermans, & D. Vansteenwegen, *Fear and learning: Basic science to clinical application* (pp. 117-132). Washington DC: American Psychiatric Association.

Luque, F. C. & O'Hora, D. (2016). Verbal reports in the experimental analysis of behavior. *Revista International Journal of Psychology & Psychological Therapy*, *16*(2), 157-177.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: Freeman.

Meulders, A., Franssen, M., Fonteyne, R., & Vlaeyen, J. W. (2016). Acquisition and extinction of operant pain-related avoidance behavior using a 3 degrees-of-freedom robotic arm. *Pain*, *157*(5), 1094-1104.

Milad, M. R., & Quirk, G. J. (2002). Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature, 420*, 70-74.

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: ten years of progress. *Annual Review of Psychology, 63*, 129-151.

Moore, J. (2002). Some thoughts on the relation between behavior analysis and behavioral neuroscience. *The Psychological Record, 52,* 261-279.

Moors, A. (2007). Can cognitive methods be used to study the unique aspect of emotion: An appraisal theorist's answer. *Cognition and Emotion, 21*, 1238-1269.

Moors, A. (2009). Theories of emotion causation: A review. *Cognition & Emotion, 23*, 625-662.

Mowrer, O. H. (1939). A stimulus-response analysis of anxiety and its role as a reinforcing agent. *Psychological Review, 46*, 553.

O'Hare, J. K., Ade, K. K., Sukharnikova, T., Van Hooser, S. D., Palmeri, M. L., Yin, H. H., & Calakos, N. (2016). Pathway-specific striatal substrates for habitual behavior. *Neuron*, *89*(3), 472-479.

O'Hora, D., Barnes□Holmes, D., & Stewart, I. (2014). Antecedent and consequential control of derived instruction□following. *Journal of the Experimental Analysis of Behavior*, *102*(1), 66-85.

Onat, S., & Büchel, C. (2015). The neuronal basis of fear generalization in humans. *Nature Neuroscience, 18*, 1811-1818.

Ostlund, S. B., & Balleine, B. W. (2009). On habits and addiction: An associative analysis of compulsive drug seeking. *Drug Discovery Today: Disease Models*, *5*(4), 235-245.

Pavlov, I. P. (1927/2010). *Conditioned re exes. An Investigation of the physiological activity of the cerebral cortex*. *Annals of neurosciences*, *17*(3).

Quirk, G. J., Garcia, R., & González-Lima, F. (2006). Prefrontal mechanisms in extinction of conditioned fear. *Biological Psychiatry, 60*, 337-343.

Quirk, G. J., Russo, G. K., Barron, J. L., & Lebron, K. (2000). The role of ventromedial prefrontal cortex in the recovery of extinguished fear. *The Journal of Neuroscience, 20*, 6225-6231.

Redgrave, P., Rodriguez, M., Smith, Y., Rodriguez-Oroz, M. C., Lehericy, S., Bergman, H., … Obeso, J. A. (2010). Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nature Reviews Neuroscience*, *11*(11), 760-772.

Roche, B. T., Kanter, J. W., Brown, K. R., Dymond, S., & Fogarty, C. (2008). A comparison of" direct" versus" derived" extinction of avoidance responding. *The Psychological Record*, 58, 443-464.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis*. BF Skinner Foundation.

Skinner, B. F. (1974). *About behaviorism*. Vintage.

Sotres-Bayon, F., & Quirk, G. J. (2010). Prefrontal control of fear: more than just extinction. *Current Opinion in Neurobiology, 20*, 231-235.

Vahey, N., & Whelan, R. (2016). The functional□cognitive framework as a tool for accelerating progress in cognitive neuroscience: On the benefits of bridging rather than reducing levels of analyses. *International Journal of Psychology*, *51*(1), 45-49.

Vahey, N. A., Nicholson, E., & Barnes-Holmes, D. (2015). A meta-analysis of criterion effects for the Implicit Relational Assessment Procedure (IRAP) in the clinical domain. *Journal of Behavior Therapy and Experimental Psychiatry*, *48*, 59-65.

Vansteenwegen, D., Vervliet, B., Iberico, C., Baeyens, F., Van den Bergh, O., & Hermans, D. (2007). The repeated confrontation with videotapes of spiders in multiple contexts attenuates renewal of fear in spider-anxious students. *Behaviour Research and Therapy, 45*, 1169-1179.

Vervliet, B., & Indekeu, E. (2015). Low-Cost Avoidance Behaviors are Resistant to Fear Extinction in Humans. *Frontiers in Behavioral Neuroscience*, *9*, 351. http://doi.org/10.3389/fnbeh.2015.00351

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of human fear acquisition and extinction within a novel arbitrary stimulus category. *PloS one, 9*, e96569.

Weiner, H. (1963). Response cost and the aversive control of human operant behavior. *Journal of the Experimental Analysis of Behavior, 6*, 415-421.

Whelan, R., Conrod, P. J., Poline, J. B., Lourdusamy, A., Banaschewski, T., Barker, G. J.,…Fauth-Bühler, M. (2012). Adolescent impulsivity phenotypes characterized by distinct brain networks. *Nature Neuroscience*, *15*(6), 920-925.

Wilensky, A. E., Schafe, G. E., Kristensen, M. P., & LeDoux, J. E. (2006). Rethinking the fear circuit: the central nucleus of the amygdala is required for the acquisition, consolidation, and expression of Pavlovian fear conditioning. *The Journal of Neuroscience, 26*, 12387-12396.

Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, *67*, 289-314.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464-476.

Yule, W., Bolton, D., Udwin, O., Boyle, S., O'Ryan, D., & Nurrish, J. (2000). The long-term psychological effects of a disaster experienced in adolescence: I: The incidence and course of PTSD. *Journal of Child Psychology and Psychiatry, 41*, 503-511.

Zentall, T. R. (2012). The heuristic value of cognitive terminology. *The Psychological Record*, 62, 321