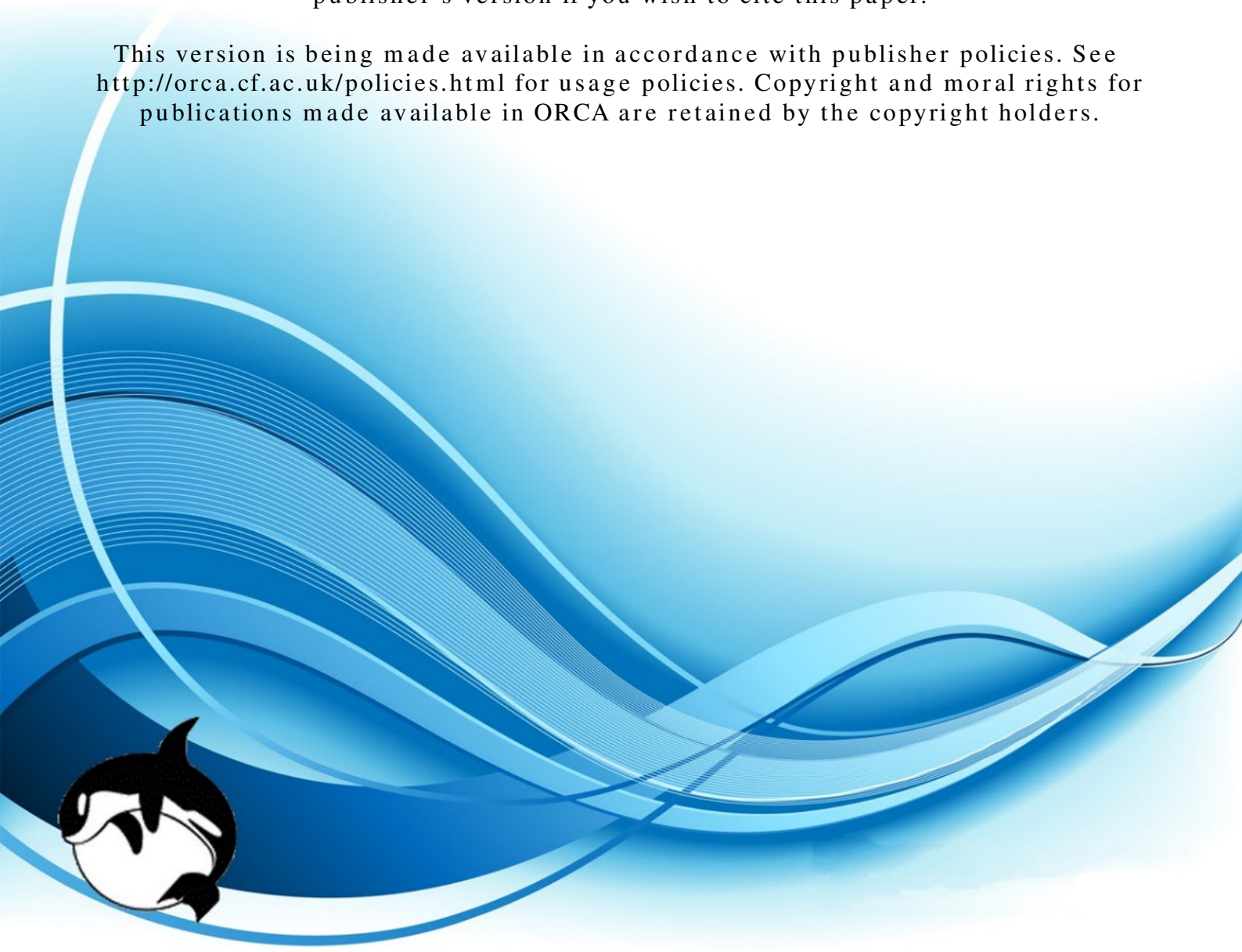

Higher-order conditioning:
A critical review and computational model

Robert C. Honey and Dominic M. Dwyer
Cardiff University

Short title: Higher-order conditioning

Accepted version February 2022: *Psychological Review*

Acknowledgements

Both authors contributed to the ideas presented in this article and to its preparation for publication. The development of the ideas, together with our research that informed them, was supported by the BBSRC (UK; BB/T004339/1; PI: RCH). A general purpose HeiDI open-source app (https://victor-navarro.shinyapps.io/heidi_app/) is available to supplement the simulations presented in this paper, which are also available from RCH in the form of annotated Excel spreadsheets. The empirical work that underpinned the ideas has been published in a series of papers, but those papers did not contain reference to the ideas formally developed here, for which this article is the original and cited source. We thank A.F. Iliescu and V.M. Navarro for their comments on a draft of this paper. Address for correspondence: Robert C. Honey, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK; Tel: +44 (0)29 20875868; Email: Honey@cardiff.ac.uk

Abstract

Higher-order conditioning results from a simple training procedure: Pairing two relatively neutral conditioned stimuli, A and X, allows properties separately conditioned to X (e.g., through pairing it with an unconditioned stimulus, US) to be evident during A. The phenomenon extends the range of ways in which Pavlovian conditioned responding can be expressed and increases its translational relevance. Given this relevance and the wealth of available behavioral analysis, it is a surprisingly underdeveloped territory for formal theoretical analysis. Here, we first provide a critical review of two (informal) classes of account for higher-order conditioning that reflect either: (1) processes that are analogous to Pavlovian conditioning, but involving associatively activated representations (e.g., A→US); or (2) the formation of an associative chain (e.g., A→X, and X→US). Our review first identifies fundamental theoretical and empirical challenges to both classes of account. We then develop a new computational model of higher-order conditioning that meets the challenges identified by showing: how reciprocal associations between A, X and the US are formed and affect performance; and how the similarity of stimuli, their traces and associatively retrieved representations modulate this process. The model generates a wealth of novel predictions, providing a platform for further empirical and theoretical analysis.

Keywords: Association; Behavior; Pavlovian conditioning; Similarity, Timing.

Introduction

The *association of ideas* is central to the philosophical roots of psychology (e.g., James, 1890; Warren, 1922), and remains a core explanatory principle and influence across psychology (e.g., Mackintosh, 1983; Rumelhart, Hinton, & Williams, 1986; Shanks, 1995), neuroscience (e.g., Schultz, Dayan, & Montague, 1997; Wimmer & Shohamy, 2012), and artificial intelligence (e.g., Grossberg, 1988; Sutton & Barto, 1981). The empirical analysis of associative learning originates in the study of Pavlovian or classical conditioning, where a conditioned stimulus (CS) comes to elicit a conditioned response (CR) as a result of being paired with an unconditioned stimulus (US; Pavlov, 1927). The fact that this CR often resembles the unconditioned response (UR) elicited by the US (e.g., Jenkins & Moore, 1973) suggests – alongside other compelling evidence – that the presentation of the CS has evoked the representation of the US through an association formed between their central representations. By this account, a seemingly reflexive CR is based on the capacity of the CS to activate the representation or idea of the US. In fact, Pavlovian conditioning continues to represent an important testbed for the study of associative processes in both non-human animals and humans (see Mackintosh, 1994; see also, for example, Gallistel & Gibbon, 2000; Honey, Dwyer & Iliescu, 2020a; Stout & Miller, 2007).

Sensory preconditioning		
A→X	X→US	A?
Second-order conditioning		
X→US	A→X	A?

Table 1. A and X denote different (e.g., auditory and visual) stimuli, and US denotes an unconditioned stimulus (e.g., an appetitive stimulus like food or an aversive stimulus like a mild shock). The critical trials depicted above are embedded in within-subjects designs or between-subject designs showing that higher-order conditioning is a consequence of the X→US and A→X pairings.

Higher-order conditioning refers to the observation that once a conditioned property has been established to one CS (e.g., an excitatory or inhibitory association with a US), other stimuli that have been paired with that CS gain some of its properties (see Rescorla, 1976). For example, if an excitatory association has been established between one CS (X) and a US, then a second stimulus (A) will come to elicit conditioned responding as a consequence of being paired with X. This effect was first identified by Pavlov and his colleagues during studies of salivary conditioning in dogs, and referred to as "*a reflex of the second order*" (Pavlov, 1927, pp. 104-106). In fact, he reported that the effect was "*in most cases very weak*". However, it has actually proven to be a marked and reliable effect across many species using the two canonical higher-order conditioning procedures: Sensory preconditioning (e.g., Brogden, 1939; Lin & Honey, 2011; Lin, Dumigan, Dwyer, Good & Honey, 2013; Rescorla & Cunningham, 1978; Rhodes, Creighton, Killcross, Good & Honey, 2009; Ward-Robinson & Hall, 1996) and second-order conditioning (e.g., Crawford & Domjan, 1995; Holland & Rescorla, 1975; Leyland, 1977; Lin & Honey, 2011; Rashotte, Griffin & Sisk, 1977; Rizley & Rescorla, 1972; Ward-Robinson, 2004). As we shall show, these forms of higher-order conditioning have an applied or translational significance that at least matches Pavlovian conditioning.

Both sensory preconditioning and second-order conditioning involve a stage of training in which two neutral stimuli (A and X) are paired, and another stage in which a conditioned property is established to X (e.g., by pairing it with a US; see Table 1). For sensory preconditioning, A→X pairings precede X→US pairings, and responding to A measured in a later test, whereas for second-order conditioning, X→US pairings precede A→X pairings, and the development of responding to A is measured. The fact that both procedures can bring about marked and reliable changes in responding to A, extends the ways in which Pavlovian conditioning can influence behavior to a broader range of real-

world settings, where events with primary motivational significance (potential USs) are relatively rare (e.g., Flagel, Clark et al., 2011; Nasser, Chen, Fiscella, & Calu, 2015; Robinson & Flagel, 2009). The two procedures also continue to provide a basis for both translational research (e.g., Wessa & Flor, 2007; see also, Field, 2006; Haselgrove & Hogarth, 2011) and neurobiological analyses of learning and memory (for a review, see Gewirtz & Davis, 2000; see also, e.g., Gilboa, Sekeres, Moskovitch & Winocur, 2014; Holland, 2016; Iordanova, Good & Honey, 2011; Lay, Westbrook, Glanzman & Holmes, 2018; Lin & Honey, 2011; Lin, Dumigan, Good & Honey, 2016; Maes, Sharpe et al., 2020; Mollick, Hazy et al., 2020; Ward-Robinson, Coutureau, Good, Honey, Killcross, & Oswald, 2001). One recent example serves to illustrate the potential of higher-order conditioning procedures to enhance our understanding of the neurobiological basis of learning, and of the role of prediction error in particular.

Maes et al. (2020) conducted an ingenious series of experiments to elucidate how dopamine transients in the VTA modulate associative learning. In one experiment, rats received pairings of two visual cues (on A→X trials) and then received compound trials where A was separately presented with two auditory stimuli (C and D) and the resulting compounds were both paired with X (i.e., AC→X and AD→X trials). If formal models of Pavlovian learning (e.g., Rescorla & Wagner, 1972) also apply to sensory preconditioning, then this arrangement should result in A blocking the development of the C→X and D→X associations (cf. Kamin, 1969): because A already predicts X and there is no prediction error to generate further learning. On AC→X trials, however, Maes et al. (2020) temporarily activated VTA dopamine neurons at the start of X to assess the role of dopamine transients in associative learning. X was later paired with food, which resulted in C being more likely than D to elicit the conditioned response that was measured, approaching the site of food delivery. Alongside other controls, these results

suggest that dopamine transients can augment associative learning involving neutral stimuli, counteracting the fact that the error signal (on AC→X trials) had been reduced by prior A→X training trials. While these results are clearly important from a neurobiological standpoint, they actually beg the question: What was learnt on AC→X and X→US trials that allowed subsequent presentations of C to elicit conditioned behavior?

Given the broad significance of higher-order conditioning – together with the wealth of relevant behavioral findings – it is a surprisingly underdeveloped territory for formal theoretical analysis. Indeed, even the informal associative accounts of higher-order conditioning that are routinely adopted and contrasted (across different levels of analysis) have not materially changed since the analysis of the phenomenon provided by Mackintosh (1974; pp. 85-91; cf. Gewirtz & Davis, 2000). Our critical review begins by describing these accounts, which have also been widely adopted across studies of human learning in a variety of contexts (*behavioral*: e.g., Craddock, Wasserman, Polack, Kosinski, Renaux & Miller, 2018; Ecker & Bar-Anan, 2019; *translational*: e.g., Davey & Arulampalan, 1982; Davey & McKenna, 1983; *neuroscientific*: e.g., Seymour, Doherty et al., 2004; Wimmer & Shohamy, 2012; Yu, Lang, Birbaumer & Kotchoubey, 2014). We proceed by identifying two key challenges to these analyses: They are inconsistent with formal models of associative learning (e.g., Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981), and fail to explain the specific conditions under which higher-order conditioning is observed. We then identify a recent formal model of Pavlovian learning and performance (HeiDI; Honey et al., 2020a) that provides a foundation for a new computational model of higher-order conditioning. The model specifies the learning rules, associative structures, and performance rules for higher-order conditioning. The final model also incorporates a function that captures the similarity (in terms of perceived intensity) of a directly activated CS representation to (1)

its less intense decaying trace, and (2) associatively activated representation. The latter function enables important features of higher-order conditioning to be explained, while providing a basis for an associative analysis of timing phenomena (see also, Staddon, 2005; Staddon & Higa, 1999).

Representation-mediated learning and associative chains

One account of higher-order conditioning assumes that it reflects the operation of processes that are analogous to those that underpin Pavlovian conditioning: The formation of an association between A and the US (cf. Pavlov, 1927, p. 105). In Pavlovian conditioning an association is held to develop between the directly activated CS representation and either the directly activated US representation (i.e., a stimulus-stimulus association) or the processes responsible for the generation of the CR (i.e., a stimulus-response association). In higher-order conditioning, however, the directly activated representations of A and the US (and the UR) occur on separate trials: A on A→X trials and the US on X→US trials. So, how could an A→US or A→UR association form? The idea is that the formation of the A→US (or A→UR) association is based on representation-mediated learning (Hall, 1996; Holland, 1981, 1983; see also, Cohen-Hatton, Haddon, George, & Honey, 2013; Craddock et al., 2018; Honey & Hall, 1991; Iordanova et al., 2011; Lin, Dumigan, Recio & Honey, 2017; Ward-Robinson, 2004). In sensory preconditioning, if the A→X trials allow the presentation of X to activate a representation of A, then the X→US trials could allow an association to develop between the associatively activated representation of A and the US representation (or processes more directly responsible for the UR; see Ward-Robinson & Hall, 1996, 1998). In second-order conditioning, if the X→US trials allow X to activate a representation of the US on the later A→X trials, then this might result in A becoming linked to the representation of

the US (Konorski, 1948, p. 68) or to a component of the process that generates the UR (e.g., Rizley & Rescorla, 1972; cf. Pavlov, 1927, p.105).

The account of higher-order conditioning, based on representation-mediated learning, is often contrasted with the idea that higher-order conditioning depends on the formation of a directional associative chain: A→X→US or A→X→UR (e.g., Gewirtz & Davis, 2000). In this case, if X→US pairings result in the formation of a (directional) association between representations of X and the US (or UR), and A→X pairings result in the formation of an association between A and X, then the tendency for A to provoke conditioned responding reflects the efficacy of the associative chain: A→X→US or A→X→UR.

Theoretical challenges

Representation-mediated learning is clearly an appealing explanation for higher-order conditioning (Hall, 1996; Holland, 1981); but the specific claim that a retrieved representation (e.g., A) can become linked by an excitatory association to a representation of a stimulus that is being directly activated (e.g., a US) is controversial (Wagner, 1981; see also, Dickinson & Burke, 1996; Van Hamme & Wasserman, 1994). In more formal terms, for the strengthening of an A→US association to occur during the X→US trials of a sensory preconditioning procedure, the learning rate parameter for A (α_A; within computational models of Pavlovian learning) would need to be positive as opposed to being zero (e.g., Rescorla & Wagner, 1972; but see, Van Hamme & Wasserman, 1994). For example, according to the Rescorla and Wagner (1972) model, the product of the learning rate parameters for a CS (α) and US (β) affect the rate of change in the CS→US association (i.e., $\Delta V_{CS\rightarrow US;}$ the subscript denotes the direction of the association) according to the following equation: $\Delta V_{CS\rightarrow US} = \alpha.\beta.(\lambda - \sum V_{TOTAL\text{-}US})$. In this equation, λ is the maximum associative strength supportable by the US, and $\sum V_{TOTAL\text{-}}$

$_{US}$ is the sum of the associative strengths of stimuli presented on the trial. If α_A takes a value of 0 when A is physically absent then mediated learning could not occur because the error term is multiplied by the product of α and β, which are aligned to the intensities of the CS and US, respectively. But even if the α_A for a retrieved A is assumed to be greater than 0, then what exact value should it take?

One could imagine that just as the value of α_X for a stimulus (i.e., X) is related to its intensity, the α_A value for a retrieved stimulus (A) is related to the strength with which it is retrieved by stimuli that are present (i.e., $\sum V_{TOTAL\text{-}A}$). Indeed, in order to generate effects associated with the nonreinforcement of a CS (e.g., extinction), Rescorla and Wagner (1972; Wagner & Rescorla, 1972) had to make a series of assumptions, including the idea that when a US is predicted but absent it has a learning rate parameter (β_I) that is positive. Otherwise, the fact that the error term (i.e., $\lambda - \sum V_{TOTAL\text{-}US}$) is negative on a trial when a previously reinforced CS is presented for extinction would result in no change in the associative strength of that stimulus. The general idea that the learning rate parameter associated with an absent stimulus is not 0 has a clear precedent. However, in the context of higher-order conditioning, the basis for the exact value of the retrieved α for A in sensory preconditioning (or β for second-order conditioning) has not been specified: It would seem peculiar to simply substitute the corresponding α and β values for when corresponding stimuli were present (i.e., irrespective of whether they are being weakly or strongly associatively activated). We have presented one simple alternative above (i.e., using a value aligned to $\sum V_{TOTAL\text{-}A}$ in place of α_A), which could be combined with a given learning rule. But, it should be noted that this possibility carries with it further complexities. For example, including retrieved αs (like adding more stimuli) increases the likelihood that the sum of αs on trial with multiple CSs will exceed 2. This results in

changes in associative strength that are no longer error correcting, but rather increasingly oscillatory; which is an underappreciated feature of the models of the form proposed by Rescorla and Wagner (1972; see McLaren & Mackintosh, 2001; p. 216). However, we will return to the complementary idea that the strength with which the memory of a stimulus is associatively activated can be a useful proxy for the (original) intensity of that stimulus, especially if the learning rule is one in which stimulus intensity sets the asymptotic value for associative strength (cf. Honey et al., 2020a).

Finally, during the A→X trials of a second-order conditioning procedure the representation of the US is being activated by X (i.e., $\sum V_{TOTAL\text{-}US} > 0$) but the US is absent (cf. Konorski, 1948, p. 68). Under these conditions, formal models of learning predict that conditioned inhibition, rather than excitation, will develop between A and the US (e.g., Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981). This prediction has been amply confirmed: Intermixing reinforced X trials with nonreinforced AX trials (the same trials types used in second-order conditioning) results in A becoming a conditioned inhibitor: capable of suppressing conditioned responding to X (and other stimuli that have been paired with the same US) and only coming to generate conditioned responding slowly when subsequently paired with the US (for a review, see Rescorla, 1969). Informal accounts of higher-order conditioning (in terms of representation-mediated learning or indeed associative chains) and formal models of Pavlovian learning have not been reconciled with the co-existence of these two well-established empirical observations: second-order conditioning and conditioned inhibition. We return to the important (theoretical and empirical) issue of when conditioned inhibition rather than second-order conditioning is observed after our formal model has been developed. First we consider some specific empirical challenges to the two informal accounts described above, which concern the conditions under which higher-order conditioning is observed

and how it is evident in behavior. These challenges inform features of the formal model that we then present.

Empirical challenges

When higher-order conditioning is observed. While higher-order conditioning itself is well established, it occurs under conditions that are a challenge to both of the informal accounts that we have considered. First, introducing a trace interval between X and the US (a procedure that reduces conditioned responding to X) can enhance sensory preconditioning to A (Lin & Honey, 2011; Ward-Robinson & Hall, 1998; see also, Kamil, 1969) and second-order conditioning to A (Lin & Honey, 2011; see also, Barnet & Miller, 1996; Cole, Barnet & Miller, 1995). If higher-order conditioning reflects representation-mediated learning or the operation of an associative chain, then without further assumptions trace conditioning should reduce higher-order conditioning. Second, extinguishing first-order conditioning to X does not (always) reduce higher-order conditioning to A in both sensory preconditioning (Ward-Robinson & Hall, 1996) and second-order conditioning procedures (e.g., Amiro & Bitterman, 1980; Archer & Sjödén, 1982; Cheatle & Rudy, 1978; Nairne & Rescorla, 1981; Rizley & Rescorla, 1972; see also, Craddock et al., 2018; but see, Rescorla, 1982). Both accounts are undermined by this observation: extinguishing X should allow mediated extinction of A (which is associatively activated, but not reinforced; see Holland & Forbes, 1982), and changing the efficacy of any link in the chain (e.g., X→US or X→UR) should also be reflected in the capacity of A to elicit responding. Third, if A is presented with X during the test for sensory preconditioning, the resulting compound generates more responding than when X is presented alone or with a control stimulus (Lin et al., 2013; Ward-Robinson et al., 2001). This finding suggests that A has a capacity to generate conditioned responding that is independent of the capacity of X to generate responding. However, as we shall show,

this capacity need not derive from representation-mediated learning (e.g., involving A and the US).

The results outlined in the previous paragraph are clearly problematic for accounts based on representation-mediated learning and standard associative chains. However, it is possible to explain the results by specifying the effective components in the associative chain (e.g., A→X→US) in more detail. First assume that the intensity of a given stimulus X is represented (e.g., α_X) and that upon presentation of X this value is high, but that it declines to a lower value after the offset of X. This will mean that the intensity of X th.at acquires associative strength during standard conditioning with X will be higher than during trace conditioning. If the efficacy of the associative chain (e.g., A→X→US) is determined by the extent to which the intensity of the representation of X that is retrieved by A (i.e., $\alpha_{X\text{-}R}$) is similar (S) to the intensity of X (α_X) during X→US trials, then higher-order conditioning could be enhanced by trace conditioning with X (see Lin & Honey, 2011, p. 321-322; see also Hull, 1943, Barnet & Miller, 1996; Cole et al., 1995; Hoffeld, Kendall, Thompson, & Brogden, 1960; Kamil, 1969; Maes et al., 2020; Ward-Robinson & Hall, 1998).

The simple idea outlined in the immediately preceding paragraph could also provide an analysis for why extinguishing X does not (always) result in a reduction in responding to A (see Amiro & Bitterman, 1980; Archer & Sjödén, 1982; Cheatle & Rudy, 1978; Craddock et al., 2018; Nairne & Rescorla, 1981; Rescorla, 1982; Rizley & Rescorla, 1972; Ward-Robinson & Hall, 1996). Given the general observation that a more salient stimulus will overshadow a less salient one (cf. Kamin, 1969; Mackintosh, 1976), when X is presented for extinction the more intense X rather than its less intense trace would undergo greater extinction. To the extent that higher-order conditioning is supported by the associatively retrieved X (i.e., $\alpha_{X\text{-}R}$) and this retrieved X is similar (in intensity) to the

memory trace of X, then extinction of X might well be ineffective in reducing responding to A. Finally, the suggestion that the memory of X retrieved by A at test can be more or less similar (in terms of intensity) to the intensity of X when it was paired with the US, provides a potential account for the observation that AX generates more conditioned responding than X (cf. Lin et al., 2013; Ward-Robinson et al., 2001): The intensity of the associatively generated X (on AX trials) might be more similar to the intensity of X on conditioning trials than is the intensity of X (on X alone trials).

The overarching idea that has been entertained above is that animals represent the similarity of (or difference between) a representation that is being associatively activated (e.g., $\alpha_{X\text{-}R}$) and one that has been directly activated (e.g., α_X); and that they do this in terms of stimulus intensity. This idea is developed more formally below in the context of our novel analysis of higher-order conditioning, where we introduce a similarity function (i.e., $\alpha_{X\text{-}R}S\alpha_X$). But, first we summarize some further evidence that is beyond the scope of traditional accounts based on representation-mediated learning and associative chains.

How higher-order conditioning is evident in behavior. The two accounts of higher-order conditioning considered thus far both assume that the critical associations that underlie performance are directional, whether the stimuli have been presented simultaneously or sequentially. In the case of representation-mediated learning, performance is assumed to be based on an association from a representation of A to the US (i.e., A→US), while the critical links in the associative chain are from A to X (A→X) and from X to the US (X→US). We could make these claims more formal and assume that the associative strength (V) of each associative link ($V_{A\rightarrow US}$, $V_{A\rightarrow X}$ and $V_{X\rightarrow US}$) can take values from 0 and 1 (e.g., depending on the intensities of A, X and the US, which also take values from 0 and 1). The efficacy of an associative chain (e.g., $V_{A\rightarrow X\rightarrow US}$) upon

presentation of A could then be a product of the two associations (i.e., the numerical value of $V_{A\rightarrow X} \times V_{X\rightarrow US}$). The simplifying assumption could then be made that the vigor or frequency of higher-order conditioned responding is ordinally related to $V_{A\rightarrow US}$ in the case of representation-mediated learning, and $V_{A\rightarrow X\rightarrow US}$ in the case of the associative chain. This assumption was made by Rescorla and Wagner (1972) in the context of how the strength of a CS→US association (i.e., $V_{CS\rightarrow US}$) relates to the vigor or frequency of conditioned responding. In any case, the assumption that the associations are directional carries with it two predictions that we know to be inaccurate.

The first prediction is a general one: Directional associations terminating in the US (whether they are direct, mediated or chained) provide a clear-cut basis for generating conditioned responses (e.g., to A and X) that reflect that nature of the US, but fail to predict that conditioned responses also reflect the nature of A and X (e.g., Holland, 1977; Patitucci, Nelson, Dwyer and Honey, 2016; Timberlake & Grant, 1975). For example, when hungry rats are given pairings of a wooden block with food they come to orient towards the block; but if they receive pairings of another rat with the delivery of food they also exhibit social behaviors toward their conspecific (Timberlake & Grant, 1975). Importantly, the dissociation of response types between the wooden block and conspecific CSs only emerged across training and was not evident in controls where the CS and US were unpaired, and so cannot be attributed simply to the conspecific eliciting social behaviour. This observation, together with those from standard conditioning procedures (e.g., Holland, 1977; Patitucci, Nelson, Dwyer and Honey, 2016), shows that the nature of both the CS and the US determines how learning is evident in behavior.

The second, related prediction directly concerns higher-order conditioning and follows from both of the informal accounts that have been the focus of interest so far: Conditioned responses established by the Pavlovian conditioning trials (e.g., X→US)

should be equivalent to those engendered by higher-order conditioning trials (e.g., A→X; cf. Pavlov, 1927). This prediction derives from the fact that higher-order conditioned behavior is generated – in one way or another – through activation of the US representation (see Holland & Rescorla, 1975). However, when multiple measures of the conditioned behaviors elicited by A and X have been taken, the conditioned responses to them turn out to be far from equivalent.

One striking example of sensory preconditioning in rodents was originally reported by Rescorla and Cunningham (1978; see also, Fudim, 1978) and has since been replicated on many occasions. For example, Dwyer, Burgess and Honey (2012) reported a series of experiments in which they first gave thirsty rats separate access to two flavor compounds that both contained two flavors (A with X and B with Y). In this case, the components of the compounds were presented simultaneously rather than sequentially. The rats then received access to X that was paired with illness and access to Y that was not. As in the original study, during the test rats not only showed a reluctance to consume X relative to Y, but were also reluctant to consume A relative to B (see Figure 1). An interesting supplementary observation reported by Dwyer et al. (2012) was that while the aversion was also evident in the way the rats consumed X relative to Y (i.e., as a reduction in lick cluster size – indicative of reduced hedonic responses; see Dwyer, 2012), there was no comparable effect during A and B. Using a flavor-aversion learning procedure in rats, Pavlovian conditioning and sensory preconditioning were not equally evident across different response measures. Why should this be the case if sensory preconditioning is based on a mediated A→illness association, or by directional associations between A and Illness that happened to be mediated by X (i.e., A→X→illness or B→Y→water)?

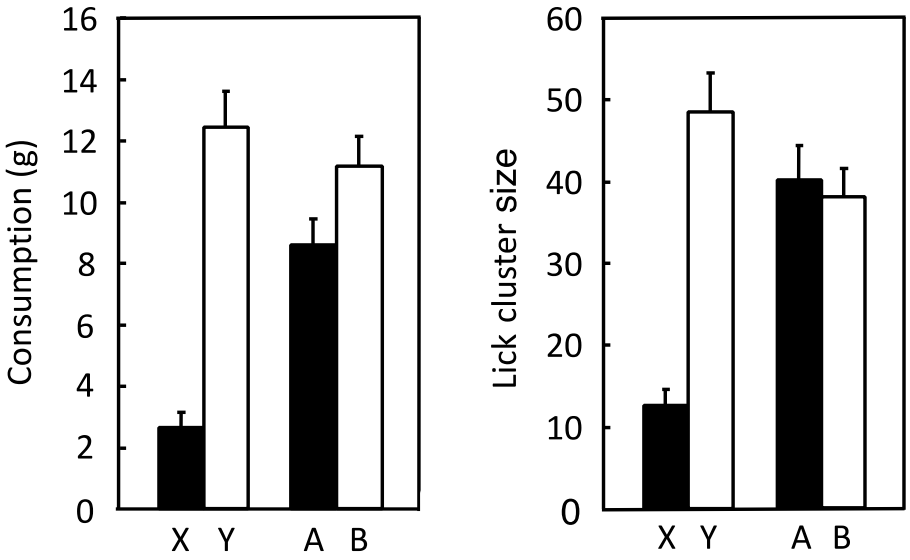

Figure 1. *Sensory preconditioning: Response measures*. Mean (+SEM) consumption of flavors X, Y, A and B (in grams; left-hand panel) and mean (+SEM) lick cluster size (right-hand panel). Prior to the test, the thirsty rats had consumed two flavor compounds (AX and BY), and then consumption of X, but not Y, was paired with the induction of illness. [Adapted from: Dwyer, D.M., Burgess, K.V., & Honey, R.C. (2012). Avoidance but not aversion following sensory-preconditioning with flavors: A challenge to stimulus substitution. *Journal of Experimental Psychology: Animal Behavior Processes*, *38,* 359-368.]

A similar pattern of results to that reported by Dwyer et al. (2012) was observed using a standard second-order conditioning procedure in pigeons (Leyland, 1977; Rashotte et al., 1977). As in the original studies, Stanhope (1992) gave pigeons, that were hungry and thirsty, an autoshaping procedure where one localized keylight (X) was paired with food and another keylight (Y) was separately paired with water (cf. Jenkins & Moore, 1973). The pigeons came to peck both keylights, but did so with greater force to the keylight paired with food (X) than that paired with water (Y). The pigeons then received second-order conditioning trials in which keylight A was paired with X and keylight B was paired with Y. The presentation of A and B both came to evoke keypecking, but these second-order keypecks did not differ in force between the two stimuli (see also, e.g., Holland, 1977; Holland & Rescorla, 1975). First-order conditioning

reflected the properties of the CSs, the pigeons pecked the localized keylights, and the properties of the USs, the force of these keypecks reflected the nature of the different reinforcers. But, second-order conditioning reflected only the fact that A and B were keylights. Why should this be the case if second-order conditioning is based on a mediated A→food or B→water associations, or by directional associations between A and food and between B and water that happen to be mediated by X and Y (i.e., A→X→food or B→Y→water)? It is simply unclear.

Higher-order conditioning is a marked and reliable phenomenon, but one that is not (or not always) apparent in the same response measures as Pavlovian conditioning: A simple observation, but one that is inconsistent with Pavlov's principle of *stimulus substitution* and with his specific analysis of second-order conditioning. He suggested that during second-order conditioning trials in which a tone was paired with a light, the tone "*has actually gone through the same process as occurred when the light received (from its association with eating) its stimulatory effect on the salivary secretion*" (Pavlov, 1927, p.105). Moreover, the related idea that higher-order conditioning simply reflects the capacity of A to activate a representation of the US (or its UR), through a mediated A→US or A→UR association or an A→X→US or A→X→UR chain, ignores a simple observation: When two neutral stimuli are paired (e.g., an auditory stimulus with a localized visual stimulus; A→X), the auditory stimulus (A) comes to generate responding that reflects the nature of the visual stimulus (X): Rats come to orient to the location in which the visual stimulus (X) is about to appear (e.g., Honey, Good & Manser, 1998; Honey, Watt & Good, 1998; Silva, Haddon et al., 2019; see also, Narbutovich & Podkopayev, 1936; cited in Konorski, 1948, p. 91). Far from being "behaviorally silent" (e.g., Dickinson, 1980, p. 5), pairing two relatively neutral stimuli can, in and of itself, result in marked changes in behavior. This observation indicates that any complete analysis of

higher-order conditioning needs to incorporate the possibility that the resulting behaviors to A will not only reflect those evoked by the US, but also those elicited by the associatively retrieved X (cf. Lin & Honey, 2011, 2016; Lin et al., 2013).

The failure to consider how the nature of X might impact behavior to A is part of a more general issue with analyses of both higher-order conditioning and Pavlovian conditioning: How do the associative structures that are formed during conditioning generate different types of behavior? The process by which behavior is generated was integral to earlier stimulus→response formulations of conditioning (e.g., Hull, 1949), but more recent formal analyses of Pavlovian conditioning (e.g., Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981) have left the process underspecified. The related (informal) accounts for higher-order conditioning that we have considered are similarly underspecified with respect to how learning is manifest in performance (e.g., Pearce, 2002). As a prelude to considering the associative structures that underpin higher-order conditioning, together with how they generate different conditioned behaviors, we first consider the following question: How does Pavlovian learning become evident in different behaviors?

Translating Pavlovian learning into behavior: HeiDI

In the case of Pavlovian conditioning, the magnitude of the US influences the form of conditioned responding. For example, increases in the magnitude of both appetitive and aversive USs result in an increase in responses that reflect the nature of the US (US-oriented CRs) and a reduction in those that reflect the nature of the CS (CS-oriented CRs; Holland, 1979; see also, Patitucci et al., 2016). These relationships between the intensity of the US and different forms of conditioned behavior are simultaneously intuitive and puzzling: If a US is intense and itself elicits a marked UR, then a CS that can associatively activate a representation of that US might be expected to generate a more marked CR

than a US that elicits a less marked UR (consistent with Pavlov's principle of stimulus substitution; Pavlov, 1927). One could also imagine that these differences would interact with CS-oriented conditioned behaviors; for example, increases in US-oriented conditioned behaviors might compete with CS-oriented behaviors (at a variety of levels). However, the critical problem for this analysis is how CS-oriented behaviors are generated by a directional link from the CS to the US representation. This problem had remained unresolved, but there is a simple solution.

Asratian (1965, pp. 150-153), reported a series of studies by M.E. Varga and Ia. M. Pressman (1958; see also, M.I. Struchkov, 1960; cited in Asratian, 1965, pp. 178-180) in which sequential presentations of two stimuli (e.g., CS→US) not only resulted in the development of conditioned responding indicative of the strengthening of an association from the CS to the US ($V_{CS \rightarrow US}$), but also responding indicative of the strengthening of the reciprocal connection between the US and CS ($V_{US \rightarrow CS}$), which included the fact that presentation of the US alone resulted in responding specific to the CS (see also, e.g., Arcediano, Escobar, & Miller, 2005; Asch & Ebenholtz, 1962; Gerolin & Matute, 1999; Tait & Saladin, 1986; Zentall, Sherburne, & Steirn, 1992). These results provided support for Pavlov's (1949, p. 452) earlier contention that the connections formed between the central processes activated by two stimuli (the "nerve points") are bidirectional or reciprocal. This contention did not take hold in the theories of Pavlovian conditioning elaborated in Western psychology, perhaps because of the apparent failure of (backward) US→CS pairings to result in US-oriented responding when the CS was presented alone (but see, for example, Cole & Miller, 1999; Heth, 1976), but probably also reflecting Pavlov's (1927) earlier emphasis on the signaling function ("signalization") of the CS (see Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981; for further discussion, see Navarro & Wasserman, 2020).

For the present purposes, however, accepting that a US→CS link is strengthened during CS→US pairings provides a potential mechanism by which standard Pavlovian conditioning results in an increase in CS-oriented responding: The CS can be assumed to have unconditioned links to response units that generate CS-oriented behaviors; and these can be amplified through the operation of the associative connections from the CS and US and critically the reciprocal connection between the US and the CS. When coupled with appropriate performance rules, the assumption that associations between the CS and US are reciprocal also provides a basis for the fact that CS-oriented and US-oriented responding are doubly dissociable: CS-oriented responding declines less rapidly over the course of extinction trials than does US-oriented responding (e.g., Iliescu, Hall, Wilkinson, Dwyer & Honey, 2018); and CS-oriented responding declines more rapidly during the presentation of a CS than does US-oriented responding (which increases; Iliescu, Dwyer & Honey, 2020). This evidence indicates that the two types of conditioned responding have distinct origins.

The formation of reciprocal associations has other desirable consequences. For example, it provides a potential explanation for why the intensity of the CS and not just that of the US (cf. Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wagner, 1981) affects the vigor of conditioned responding at asymptote (e.g., Kamin, 1965; Scavio & Gormezano, 1974). It also provides an explanation for the fact that the formation of associations between the elements of a compound (AX) can be disrupted by the presentation of the US after that compound, which we will return in the context of higher-order conditioning (Holland, 1980a; Urcelay & Miller, 2009). Briefly, the formation of a US→A (or US→X) association will limit the development of associations between A and X.

Figure 2 depicts the associative structure that forms the basis of the HeiDI model of Pavlovian conditioning (for a full discussion, see Honey et al., 2020a). It is assumed that the CS and US enter conditioning capable of generating a variety of (unconditioned) responses: The CS more likely to activate r1-r3 (CS-oriented responses) and the US more likely to activate r4-r6 (US-oriented responses). As a consequence of CS→US conditioning trials, reciprocal CS→US and US→CS associations form, according to learning rules that we will present in the context of our analysis of higher-order conditioning.

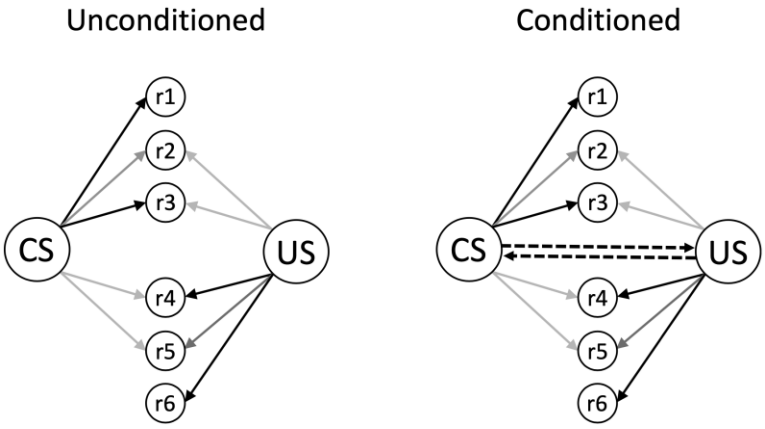

Figure 2. *A schematic associative structure for the translation of excitatory Pavlovian conditioning into performance*. The unconditioned structure on the left shows the unconditioned links from the CS and US to response-generating units (r1-r6) before conditioning, with the darkness of the arrows indicating their strength. The conditioned structure on the right shows the reciprocal associations between the CS and US nodes (denoted by the dashed lines) that are assumed to develop as a consequence of CS→US pairings. [Adapted from: Honey, R.C., Dwyer, D.M., & Iliescu, A.F. (2020). HeiDI: A model for Pavlovian learning and performance with reciprocal associations. *Psychological Review*, *127*, 829-852.]

HeiDI separates the strength of the minimal (cell) assembly (Hebb, 1949) created by the CS→US and US→CS associations, from how the relative intensities of the CS and US affect performance (Hull, 1949). Thus, upon presentation of the CS, the strengths of the CS→US and US→CS associations are combined into a value ($V_{COMB\ CS⇄US}$; ⇄ denoting the combination) in a way that reflects the fact that while the CS→US association is being directly activated by the CS, the US→CS is not. More specifically, $V_{COMB\ CS⇄US}$

is equal to $V_{CS\rightarrow US}$ plus (the numerical value of) $V_{CS\rightarrow US}$ multiplied by $V_{US\rightarrow CS}$. This combined value is then distributed into CS-oriented and US-oriented components (R_{CS} and R_{US}, respectively) in proportion to the relative (perceived) intensities of the CS (α_{CS}) and the retrieved US (β_{US}, aligned to $V_{CS\rightarrow US}$; see Holland, 1977, 1979; Patitucci et al., 2016): When α_{CS} is high relative to β_{US}, the CS-oriented component tends to dominate the US-oriented component, and when β_{US} is high relative to α_{CS} the reverse is true. We now use the principles from the HeiDI model of Pavlovian conditioning as the foundation for explaining the phenomenon of interest here: higher-order conditioning. Our analysis is built on the general idea that when A is presented at test its capacity to generate different behaviors is based on any direct (reciprocal) associations that A has with the US, and the capacity of A to associatively activate the assembly involving X and the US (i.e., $V_{COMB\ X\rightleftarrows US}$).

Associative structures for higher-order conditioning

We assume that A→X and X→US trials result in reciprocal associations between the representations of A and X (i.e., A→X and X→A), and between X and the US (i.e., X→US and US→X). Figure 3 follows the format of Figure 2 in illustrating the how these associations are embedded in an unconditioned set of links between the nodes activated by A, X and the US and a set of response-generating units (again represented by r1-r6). The suggestion that the relevant associations between A, X and the US are reciprocal enables an analysis of three additional characteristics of higher-order conditioning. Once this evidence has been presented, we specify the learning rules governing the formation of these (reciprocal) associations and how they impact performance; and finally integrate this analysis with a formal specification of how a stimulus, its trace, and retrieved representations are coded.

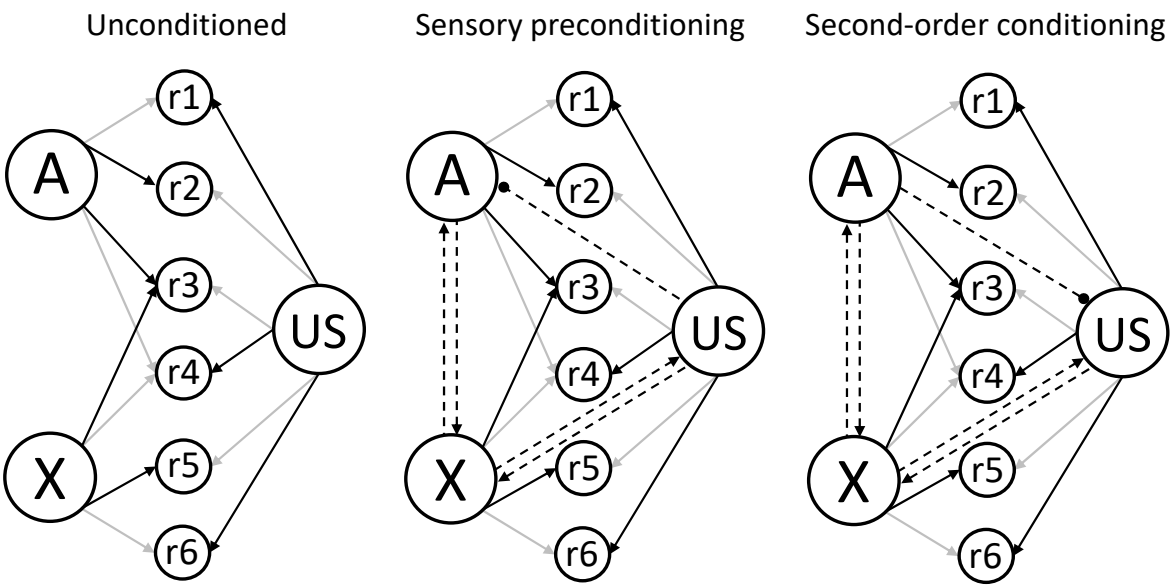

Figure 3. *Schematic associative structures for the translation of higher-order (excitatory) conditioning into performance*. The unconditioned structure on the left shows the unconditioned links from the CSs (A and X) and US to response-generating units (r1-r6) before conditioning, with the darkness of the arrows indicating their strength: A is strongly linked to r2 and r3, B is strongly linked to r3 and r5; and the US is strongly linked to r1, r4 and r6 (the remaining unconditioned links are weak or absent). The conditioned structures in the middle and right show the reciprocal associations between the A and X, and between X and the US nodes (denoted by the dashed lines with arrow heads) that develop as a consequence of higher-order conditioning trials (e.g., A→X and X→US). In the case of sensory preconditioning (center panel), there is an additional directional inhibitory US→A association (developed because A is associatively activated by X, but is absent when the US is present); whereas for second-order conditioning (right panel) there is an inhibitory A→US association (developed because the US is associatively activated by X, but is absent when A is present). Both inhibitory connections are denoted by the dashed line with the circular end; based upon one interpretation of inhibitory learning (see Honey et al., 2020a).

The assumption that the associations between A and X are reciprocal, enables an explanation for higher-order conditioning effects that have been puzzling from the perspective of standard associative chains. For example, it provides a simple explanation for backward sensory preconditioning (Ward-Robinson & Hall, 1996, 1998). Here, the first stage of the procedure involves X→A pairings rather than the more typical A→X pairings. The fact that later X→US pairings result in conditioned responding to A is taken to be inconsistent with an analysis based on a (simple) A→X→US chain because the original X→A pairings and the resulting directional X→A association would not allow the

presentation of A to activate the requisite associative chain: A→X→US. As already noted, the fact that backward sensory preconditioning is effective could indicate that when X is presented for conditioning, the associatively activated memory of A is retrieved and enters into association with the US (Ward-Robinson & Hall, 1996; see also, Dwyer, Mackintosh & Boakes, 1998). But, once it is assumed that the X→A pairings result in the formation of reciprocal associations between the representations of X and A (i.e., X→A and A→X) then the presence of the A→X→US associative chain can generate conditioned responding.

Similarly, when A→X trials are followed by US→X trials, the presentation of A provokes considerable (US-oriented) conditioned responding in spite of the fact that such backward US→X conditioning trials are a relatively ineffective way of generating US-oriented responding when X is separately tested (see Miller & Barnet, 1993). This observation has been taken to support the suggestion that the association between two stimuli includes a temporal code, and that when these temporally coded associations are superimposed, using X as the common referent, the animal should expect the US during A. However, an associative chain in which all links are reciprocal provides a plausible and simple alternative account.

Finally, in a complex set of experiments, Holland (1980a) showed that second-order conditioning to A was reduced when the US was presented on the A→X trials (i.e., A→X→US). This result is to be expected if the presentation of the US competes with A for association with X (and with X for association with A). Disrupting the formation of an association between A and X would disrupt changes in responding to A that reflected second-order conditioning. A similar analysis can be applied to results showing that when a compound of two stimuli (AX) is immediately followed by a US, A elicits little conditioned responding having been overshadowed by X, but when there is a trace interval between

AX and the US, conditioned responding to A is enhanced (Urcelay & Miller, 2009). Briefly, the association between A and X will itself be overshadowed to the extent that the US becomes associated with X, which will be less likely when there is a trace interval between AX and the US. This will allow A to activate X and to borrow its tendency to elicit conditioned responding (through the associative chain: A→X→US). This interaction, between the formation of US→X and A→X associations, has been simulated in the context of theoretically important phenomena from Pavlovian conditioning (see Honey et al., 2020ab, 2022). Another possible source for the results described by Urcelay and Miller (2009) can be derived from the suggestion that when A activates the memory of X (i.e., $\alpha_{X\text{-}R}$) at test it will be more similar to the memory of X that is conditioned when there is an interval between AX and the US than when there is no interval (cf. Lin & Honey, 2011). As noted above, we provide a formalization of this idea, after the basic learning and performance rules have been presented.

Learning rules for reciprocal associations

The generalized learning rules for Pavlovian conditioning (e.g., involving X and the US) and for associations between one CS and another (e.g., A and X) are readily specified for any two stimuli (1 and 2) with perceived intensities of α_1 and α_2: $\Delta V_{1\rightarrow 2} = \alpha_1(c.\alpha_2 - \Sigma V_{TOTAL\text{-}2})$; and $\Delta V_{2\rightarrow 1} = \alpha_2(c.\alpha_1 - \Sigma V_{TOTAL\text{-}2})$. The constant (c = 1 in units of V) is required to balance the equations in terms of the dimensions/units involved (see Honey et al., 2020a). Here, the changes in the reciprocal associations between the two stimuli presented on a given trial ($\Delta V_{1\rightarrow 2}$ and $\Delta V_{2\rightarrow 1}$) are determined by pooled error terms (($c.\alpha_2 - \Sigma V_{TOTAL\text{-}2}$) and ($c.\alpha_1 - \Sigma V_{TOTAL\text{-}1}$)). Within these error terms $\Sigma V_{TOTAL\text{-}2}$ and $\Sigma V_{TOTAL\text{-}1}$ denote the associative strength of all stimuli presented on the trial with respect to the subscripted stimulus ($_1$ or $_2$; cf. Rescorla & Wagner, 1972; see also, McLaren, Kaye & Mackintosh, 1989). Thus, both the asymptotes and the rates at which they are reached

are determined by α_1 and α_2. These generic equations are readily applied to the formation of the critical associations in higher-order conditioning experiments: reciprocal A→X and X→A associations (Equations 1 and 2, respectively) reciprocal X→US and US→X associations (Equations 3 and 4, respectively; where β_{US} sets the maximum associative strength in Equation 3 and the learning rate in Equation 4 for the US).[1]

1. $\Delta V_{A\rightarrow X} = \alpha_A(c.\alpha_X - \sum V_{TOTAL\text{-}X})$
2. $\Delta V_{X\rightarrow A} = \alpha_X(c.\alpha_A - \sum V_{TOTAL\text{-}A})$
3. $\Delta V_{X\rightarrow US} = \alpha_X(c.\beta_{US} - \sum V_{TOTAL\text{-}US})$
4. $\Delta V_{US\rightarrow X} = \beta_{US}(c.\alpha_X - \sum V_{TOTAL\text{-}X})$

These rules represent rationalizations of the Rescorla and Wagner (1972) equation, $\Delta V_{CS\rightarrow US} = \alpha\beta(\lambda - \Sigma V)$, where there is no independent free parameter lambda (λ) that determines the asymptote for the $V_{1\text{-}2}$ association (which would also be needed for the $V_{2\text{-}1}$ association). Similarly, there are no separate learning rate parameters for trials on which the target of the association is present (e.g., β_E) or absent (e.g., β_I; which would also be needed for the $V_{US\rightarrow CS}$ association; see Honey et al., 2020b). Remember that β_I

[1]A different approach to balancing the dimensions in the equations is to replace all terms in the learning and performance rules with an activation value (Act). For example, if instead of $\Delta V_{1\text{-}2}$ one specified the change in the capacity of the presentation of one stimulus (1) to produce activation in the memory of another (2; i.e., $\Delta Act_{1\text{-}2}$) then: $\Delta Act_{1\text{-}2} = Act_1 \times (Act_2 - \Sigma Act_{TOTAL\text{-}2})$; and for the reciprocal association: $\Delta Act_{2\text{-}1} = Act_2 \times (Act_1 - \Sigma Act_{TOTAL\text{-}1})$. Now, a given stimulus (e.g., 1) would accumulate the capacity to activate another (e.g., 2), and these capacities could be combined in the case of $\Sigma Act_{TOTAL\text{-}1}$ and $\Sigma Act_{TOTAL\text{-}1}$. The fact that Act values are all on the same non-dimensional scale, which can be aligned to perceived intensity, obviates the need for the constant c to address the combination of non-dimensional (α, β) with dimensional scalars (V). These activation values could also be used in the performance rules for determining the proportions of CS- and US-oriented responding (e.g., Act1/(Act1+Act2)). This approach loosens the coupling between HeiDI and a specific (associative) interpretation of Pavlovian conditioning and higher-order conditioning: It is agnostic about how one stimulus activates another (cf. Grossberg, 1980; McClelland, & Rumelhart, 1981; Wagner, 1981).

was required by the Rescorla-Wagner model because otherwise learning would not occur when the US was absent (i.e., if β = 0), and the product of the learning rate parameters ($\alpha\beta$) would therefore = 0.

The translation of higher-order conditioning into different behaviors

We have considered how the associative structure depicted in Figure 2 determines performance in Pavlovian conditioning, including the fact that the CR reflects the properties of both the CS and US (see Honey et al., 2020a). Briefly, we assumed that the combined strength of the CS→US and US→CS links ($V_{COMB\ CS\rightleftarrows US}$) influenced CS-oriented and US-oriented behaviors according to the relative perceived intensities of the CS and (retrieved) US. If higher-order conditioning involves the complex associative structures depicted in Figure 3, then how do they affect behavior? To address this question, we now extend the approach developed in the context of Pavlovian conditioning to these structures.

When A is presented, two associative structures are (potentially) important in determining higher-order conditioned responding: First, the strength of the assembly involving direct links between A and the US (i.e., $V_{COMB\ A\rightleftarrows US}$), which is calculated as before: $V_{COMB\ A\rightleftarrows US} = V_{A\rightarrow US} + (\frac{1}{c}.V_{A\rightarrow US} \times V_{US\rightarrow A})$.[2] This value will be 0 in the case of sensory preconditioning (because $V_{A\text{-}US} = 0$) and negative (i.e., inhibitory) in the case of second-order conditioning (see Figure 3). Second, the strength of the associative chain, $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$, which is calculated by multiplying the strength with which X is being activated by A by the value of $V_{COMB\ X\rightleftarrows US}$; that is, $V_{CHAIN\ A\rightarrow X\rightleftarrows US} = \frac{1}{c}.V_{A\rightarrow X} \times V_{COMB\ X\rightleftarrows US}$.

[2]Note that while multiplying a dimensionless scalar (e.g., α_A) by the constant c (1 in units of V) transforms it into units of V, multiplication of a value in units of V by the reciprocal of the constant c (i.e., $\frac{1}{c}$) returns a dimensionless value.

In this way, A can borrow the combined strength of the reciprocal associations between X and the US in both sensory preconditioning and second-order conditioning.

$$5.\quad R_A = \frac{\alpha_A}{\alpha_A + \alpha_X + \beta_{US}} (V_{CHAIN\ A \rightarrow X \rightleftarrows US} + V_{COMB\ A \rightleftarrows US})$$

$$6.\quad R_X = \frac{\alpha_X}{\alpha_A + \alpha_X + \beta_{US}} (V_{CHAIN\ A \rightarrow X \rightleftarrows US} + V_{COMB\ A \rightleftarrows US})$$

$$7.\quad R_{US} = \frac{\beta_{US}}{\alpha_A + \alpha_X + \beta_{US}} (V_{CHAIN\ A \rightarrow X \rightleftarrows US} + V_{COMB\ A \rightleftarrows US})$$

Equations 5-7 describe the way in which the relative intensities of A, X and the US (generated from the proportion terms) determine how the combined influence of the two associative structures ($V_{COMB\ A \rightleftarrows US}$ and $V_{CHAIN\ A \rightarrow X \rightleftarrows US}$ in the bracketed term) are distributed into three components, denoted R_A, R_X and R_{US}. Briefly, these components affect the response units connected to A, X, and the US. The value used in proportion terms when a stimulus is present (e.g., A) is aligned to its perceived intensity (e.g., α_A from Equations 1-4). However, when a stimulus is absent (e.g., X and the US when A is presented) their perceived intensities (α_X and β_{US}, respectively) are derived from the strengths with which they are associatively activated (i.e., by A). More specifically, when A is presented α_A is directly given, while α_X is derived from the strength with which the representation of X is directly activated by A (i.e., $|\frac{1}{c}.V_{A \rightarrow X}|$) plus the effect of any indirect link mediated by the US (i.e., $V_{A \rightarrow US \rightarrow X} = |\frac{1}{c}.V_{A \rightarrow US} \times \frac{1}{c}.V_{US \rightarrow X}|$). This mediated link will be 0 in the case of sensory preconditioning and negative in the case of second-order conditioning (see Figure 3). For sensory preconditioning, β_{US} is derived from the strength with which the US representation is activated by A via the memory of X (denoted $V_{A \rightarrow X \rightarrow US}$

= $|\frac{1}{c}.V_{A\rightarrow X} \times \frac{1}{c}.V_{X\rightarrow US}|$), while for second-order conditioning it is derived from the direct link between A and the US (i.e., $|\frac{1}{c}.V_{A\rightarrow US}|$) plus the mediated link ($V_{A\rightarrow X\rightarrow US}$).

An equivalent treatment can be applied to the presentation of X within higher-order conditioning procedures. Upon presentation of X, the bracketed term within the equivalent equations to Equations 5-7 is the sum of $V_{COMB\ X\rightleftarrows US}$ and $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$, which are both calculated in a directly analogous fashion $V_{COMB\ A\rightleftarrows US}$ and $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$. In this case, the resulting bracketed term is distributed into the R_A, R_X and R_{US} components in proportion to the perceived intensity of X (αX) and the strengths with which A (for αA) and the US (for βUS) are being (directly or indirectly) associatively activated.

The components derived from Equations 5-7 (i.e., R_A, R_X and R_{US}) are assumed to affect different behaviors through their impact on the unconditioned links between A, X and the US and r1-r6: R_A and R_X predominantly affecting CS-oriented responding (r2, r3 and r5), and R_{US} predominantly affects US-oriented responding (r1 and r6; see Figure 3). We can simply assume that upon presentation of A or X the (dimensionless) values of R_A, R_X and R_{US}, are multiplied by the vector of weights from A, X and the US to r1-r6 to determine their ultimate influence on these units.

The basic approach described above has been formally implemented and successfully applied to different phenomena involving complex interactions between associations involving CSs (e.g., A and X) and a US, on the one hand, and associations between different CSs (e.g., A and X) on the other hand (Honey et al., 2020abc, 2022). However, the application of the approach to higher-order conditioning *per se* was beyond the scope of the original formulation (see Honey et al., 2020a; p. 846).

Simulations of sensory preconditioning. The upper panels of Figure 4 depict computer simulations of sensory preconditioning in which $\alpha A = \alpha X = \beta US = .80$. The upper left-hand panel shows the values of the R_A, R_X and R_{US} components when A is assessed.

These values were calculated after 10 A→X trials and 2 X→US trials; from which point there is relatively little change in the output values for the associations between A and X or between X and the US (or the values of R_A, R_X and R_{US}). The fact that the values in Figure 4 are positive indicate that higher-order conditioning has been successfully simulated. Inspection of the left-hand panel, reveals that output values were positive and similar for R_A and R_X and both were higher than R_{US}. The fact that R_A and R_X are similar reflects the fact that they have the same α value and that $V_{A\rightarrow X}$ (which is the numerator in Equation 6) ≈ α_X as a result of approaching asymptote over the 10 A→X trials. R_{US} has a lower value, in spite of the fact that $\alpha_A = \alpha_X = \beta_{US} = .80$, because the numerator in Equation 7 is derived from the absolute numerical value of $V_{A\rightarrow X}$ multiplied by $V_{X\rightarrow US}$ (i.e., it is aligned to the perceived intensity of the US retrieved by A via X). Further simulations demonstrated that the magnitude of sensory preconditioning decreases when α_A and α_X are set to lower values in Equations 1-4, and the direct and indirect effects of these lower values are carried through to Equations 5-7. Lowering these values also reduces the tendency for sensory preconditioning to be evident in R_A and R_X (i.e., CS_A-oriented and CS_X-oriented behaviors) rather than R_{US} (i.e., US-oriented behaviors).

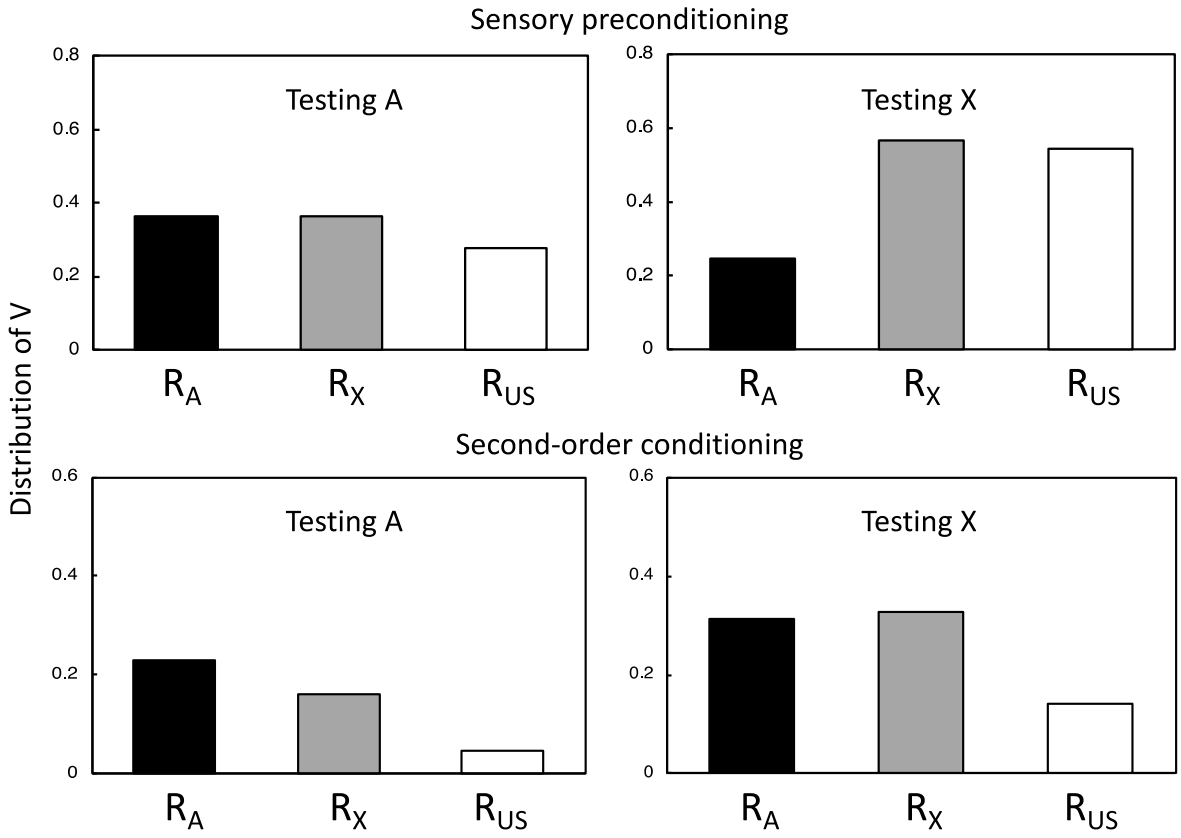

Figure 4. Computer simulations of sensory preconditioning and second-order conditioning. The output values for R_A, R_X and R_{US} were generated for A and X using Equations 1-7 together with information in the text. α_A, α_X and β_{US} were set at .80. These parameters result in similar levels of CS- and US-oriented responding in simulations of standard conditioning (see Honey et al., 2020a). There were 10 A→X trials and 2 X→US trials for the sensory preconditioning simulation, and 10 X→US trials and 2 A→X for the second-order conditioning simulation. For both simulations, the values of a R_A, R_X and R_{US} were then computed for A and X.

The upper right-hand panel of Figure 4 shows the corresponding values when X is assessed. Inspection of this panel reveals that R_A is lower than R_X (because the value of $V_{X\rightarrow A}$ declines during X→US pairings), and R_X is similar to R_{US} because α_X and β_{US} are the same (i.e., .80) and $\beta_{US} \approx V_{X\text{-}US}$. At a general level of description, these simulations reveal that while the values of R_A (that predominantly generates CS-oriented responding) are similar whether A or X is tested, R_{US} (that predominantly generates US-oriented responding) has a higher value during X than A. The results of the simulations are reminiscent of the results reported by Dwyer et al. (2012) if one equates consumption of

a fluid with CS-oriented responding (generated by R_A and R_X) and lick cluster size with US-oriented responding (generated by R_{US}).

Simulations of second-order conditioning. The lower two panels of Figure 4 show computer simulations of second-order conditioning and depict the output values for R_A, R_X and R_{US} calculated after 10 X→US trials and 2 A→X trials with the same parameter values as for sensory preconditioning. Comparison of the upper panels with the lower panels (noting the difference in scales) reveals that while the overall output values for R_A, R_X and R_{US} tended to be lower during A in second-order conditioning than in sensory preconditioning (cf. Barnet, Graham & Miller, 1991), this was especially true for R_{US}. The general reduction in R_A, R_X, and R_{US} in second-order conditioning relative to sensory preconditioning is because A rapidly acquires inhibition during second-order conditioning (cf. Rescorla & Wagner, 1972; Wagner & Rescorla, 1972), which means that the values of the bracketed terms from Equations 5-7 are smaller. Indeed, as we later demonstrate, while intermixing nonreinforced AX trials with reinforced X trials results in A acquiring (net) conditioned inhibition when α_A and α_X are low (e.g., .40), it results in more marked and protracted (net) second-order conditioning when α_A and α_X are high (e.g., .80). We know of no evidence that has examined this prediction while taking multiple measures of (CS-oriented and US-oriented) conditioned responding.

Returning to Figure 4, the fact that R_{US} takes a particularly low value during second-order conditioning (relative to R_A and R_X) reflects the effect of the inhibitory $V_{A\rightarrow US}$ on the calculated value of β_{US} for the proportion term, and the fact that $V_{X\rightarrow US}$ extinguishes during nonreinforced AX trials. More specifically, for the test with A, $\beta_{US} = |\frac{1}{c}.V_{A\rightarrow US}$ (inhibitory) + ($\frac{1}{c}.V_{A\rightarrow X}$ (excitatory) × $\frac{1}{c}.V_{X\rightarrow US}$ (excitatory))|; and for the test with X, β_{US} = $|\frac{1}{c}.V_{X\rightarrow US}$ (excitatory) + ($\frac{1}{c}.V_{X\rightarrow A}$ (excitatory) × $\frac{1}{c}.V_{A\rightarrow US}$ (inhibitory))|. Another difference with

respect to sensory preconditioning is that when A is tested the output value for R_A is higher than for R_X. This reflects the fact that the more extensive A→X trials in sensory preconditioning than second-order conditioning results in the numerator in Equation 6 (i.e., $|\frac{1}{c}.V_{A\rightarrow X}|$) being closer to the asymptote determined by α_X; but it also reflects the fact that in second-order conditioning the numerator includes the influence of $V_{A\rightarrow US\rightarrow X}$ (i.e., $\frac{1}{c}.V_{A\rightarrow US} \times \frac{1}{c}.V_{US\rightarrow X}$) which has a negative value. In any case, the pattern of results from the simulations of second-order conditioning are similar to those reported by Stanhope (1992) if one equates the level of keypecking with CS-oriented responding (generated by R_A and R_X) and the force of those keypecks with US-oriented responding (generated by R_{US}; see also, Holland, 1980b). Moreover, when the same simulations are conducted with the US present during the second stage (i.e., X→US and then A→X→US; as in a blocking procedure, Kamin, 1969), second-order conditioning to A does not occur (all values for Testing A in Figure 5 ≈ 0). This is because the US→X association (formed during X→US trials) prevents the formation of the A→X association (cf. Holland, 1980a; see also, Urcelay & Miller, 2009). Of course, the (direct) A→US association is blocked by the X→US association.[3]

The simulations of higher-order conditioning derived from Equations 1-7 show how they generate different components (R_A, R_X, and R_{US}), which support different types of conditioned behaviors once juxtaposed with the structures depicted in Figure 3. The influence of these components on the set of response units (r1-r6) with which they have

[3]Honey et al. (2020ab, 2022) present additional simulations demonstrating the role of reciprocal (CS→US and US→CS) associations in generating theoretically important phenomena in the domain of Pavlovian conditioning (e.g., downshift unblocking: Dickinson, Hall & Mackintosh, 1976; unequal associative change during compound conditioning: Rescorla, 2000; and relative validity: Wagner, Logan, Haberlandt, & Price, 1968).

(unconditioned) connections can be generated by summing the products of multiplying the numerical values of R_A, R_X and R_{US} by the strengths of the links from their corresponding nodes (i.e., A, X and the US) to each response unit. In the interests of simplicity, it can then be assumed that the resulting values (in units of V for each response unit) are reflected in the distribution of the corresponding responses (see Honey et al., 2020a). However, while these simulations confirm that a model generating the associative structures depicted in Figure 3 can provides a basis for the different forms of responding observed during higher-order conditioning, they do not address other features of the conditions under which it is observed.

The similarity of stimuli to their traces and retrieved representations

As we have just shown, Equations 5-7 capture the idea that the relative intensities of stimuli in a test pattern (e.g., A, X and the US) determine how the associative structures generated through Equations 1-4 could affect different aspects of behavior. However, these equations simply assume that the stimuli in the test pattern (in the proportion terms; e.g., X in Equation 6) and those presented during conditioning (and part of the bracketed term) are identical (e.g., in their intensity). This is clearly an oversimplification that needs to be addressed for three reasons.

1. Earlier on we introduced the general idea that animals represent intensity as part of the effective CS, in the same way that they represent other dimensions like the frequency of a pure tone. This idea formed the basis of an informal account for the fact that when there is a trace interval between X and the US, higher-order conditioning is enhanced (Lin & Honey, 2011; Ward-Robinson & Hall, 1998; see also, Barnet & Miller, 1996; Cole et al., 1995, Kamil, 1969). It was argued that if the intensity of the representation associatively retrieved by A during the test (i.e., $\alpha_{X\text{-}R}$) was more similar to the intensity of X encoded during trace conditioning than during standard conditioning,

then there would be grounds for trace conditioning enhancing higher-order conditioning to A. This informal idea clearly needs to be captured more formally in the context of a model of higher-order conditioning, but it is also required in the context of Pavlovian conditioning phenomena.

2. It is well established that animals can learn discriminations involving (a) different intensities of the same stimulus (e.g., Inman, Honey & Pearce, 2016; for a review, see Inman & Pearce, 2018), and (b) different components of the trace of the same stimulus (e.g., Lin & Honey, 2010; see also, Mackintosh, 1974; Pavlov, 1927, p. 104). Parsimony suggests that these two findings could be reduced to the operation of a single process, with traces that are more temporally removed from their stimulus source having a lower intensity (a lower α value; cf. Staddon, 2005; Staddon & Higa, 1999).

3. Finally, Equations 5-7 have no integral process for confining conditioned behavior to stimuli present on conditioning trials or those associated with them; and while the presence of associatively neutral stimuli might be expected to influence the distribution of associative strength, it seems implausible to think that they will generate anything but unconditioned responses (cf. Pavlov, 1927, p. 44; see also, Honey et al., 2020a).

We can assume that the α value of a stimulus (X) is coded on a conditioning trial and is lower not only when the physical intensity of that stimulus is reduced, but also when there is a trace interval between X and the US than when there is not. There is a clear need to identify a function that more formally specifies the similarity between two different intensities of the same stimulus, but also between an associatively retrieved stimulus (that we will simply denote as $\alpha_{X\text{-}R}$) and the intensity of the same nominal stimulus when the US is delivered (α_X). But, what is the relationship between $\alpha_{X\text{-}R}$ and α_X?

When A is presented for test in a higher-order conditioning procedure it is a simple matter to align the perceived intensity of the retrieved X (i.e., $\alpha_{X\text{-}R}$) with the numerical value of the strength of the association between A and X (i.e., $|\frac{1}{c}.V_{A\rightarrow X}|$): $V_{A\text{-}X}$ approaches the asymptote determined by α_X during A→X trials. The similarity (S) of $\alpha_{X\text{-}R}$ to α_X (i.e., $\alpha_{X\text{-}R}S\alpha_X$) can then be calculated using Equation 8, which has some simple properties. For example, $\alpha_{X\text{-}R}S\alpha_X$ approaches 1 as the values of $\alpha_{X\text{-}R}$ and α_X converge and approaches 0 as they diverge. In fact, the rate at which $\alpha_{X\text{-}R}S\alpha_X$ approaches 1 over the course of a series of (e.g., A→X) trials is invariant with respect to the target value of the association (α_X for $V_{A\rightarrow X}$; a formal proof is available on request). In contrast, the rate at which 1 is approached increases with the value of the learning rate parameter for that association (e.g., α_A for $V_{A\rightarrow X}$; see later simulations).

$$8.\ \alpha_{X\text{-}R}S\alpha_X = \frac{\alpha_{X\text{-}R}}{(\alpha_{X\text{-}R} + |\alpha_X - \alpha_{X\text{-}R}|)} \times \frac{\alpha_X}{(\alpha_X + |\alpha_X - \alpha_{X\text{-}R}|)}$$

How does $\alpha_{X\text{-}R}S\alpha_X$ affect higher-order conditioning? We assume that $\alpha_{X\text{-}R}S\alpha_X$ modulates the associative chain, $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$, within the bracketed terms of Equations 5-7, which now take the form: $(\alpha_{X\text{-}R}S\alpha_X \times V_{CHAIN\ A\rightarrow X\rightleftarrows US}) + V_{COMB\ A\rightleftarrows US}$. If A→X training resulted in $V_{A\rightarrow X}$ approaching asymptote, then the values of $\alpha_{X\text{-}R}$ and α_X would be maximally similar; assuming that α_X during these conditioning trials is the same as during A→X trials, as it normally is. But, what if α_X takes one value for A→X trials (e.g., .50) and is reduced during X→US trials (e.g., .45)? We assume that this manipulation has an equivalent effect to introducing a trace interval during X→US conditioning: where α_X would be subject to a process of trace decay before the US is delivered (cf. Lin & Honey, 2011, 2016; Lin et al., 2013). In both cases, the perceived intensity of X that gains most

associative strength (α_X) will be lower than during standard conditioning where X and the US are temporally contiguous (see Iliescu et al., 2020; see also, Holland, 1980b). Moreover, it should be clear if $V_{A\rightarrow X}$ (i.e., α_X) has not reached asymptote during A→X trials, its numerical value will more closely match .45 than it will match .50; but that as $V_{A\rightarrow X}$ tends to .50 (through increasing the number of A→X trials) the value of $V_{A\rightarrow X}$ will be closer to .50 than to .45. According to this analysis, there will be a nonmonotonic relationship between number of A→X trials and $\alpha_{X\text{-}R}S\alpha_X$ when there is a trace interval between X and the US in higher-order conditioning procedures. This relationship should result in more marked higher-order conditioning with fewer A→X trials. There is some evidence that is consistent with this prediction from studies of sensory preconditioning (e.g., Hoffeld et al., 1960; but see, Prewitt, 1967). The obvious complementary prediction, which has not been investigated, is that (physically) reducing the intensity of X between A→X trials and X→US trials will also result in a nonmonotonic relationship between the number of A→X trials and sensory preconditioning.[4]

Computer simulations confirm the accuracy of the analysis presented above. Figure 5 shows how the number of A→X training trials affects the similarity (in terms of perceived intensity) of the memory of X that is associatively retrieved by A at test (i.e., $\alpha_{X\text{-}R}$) to the memory of X at the point when the US is delivered on X→US trials (i.e.,α_X).

[4]It should be noted that $V_{COMB\ A\rightleftarrows US}$ within the bracketed term would also, in principle, be modified by the similarity between A at test and A on the A→X trials (e.g., where the inhibitory A→US link was established during second-order conditioning). But, in the case considered here $\alpha_{A\text{-}R}S\alpha_A = 1$, because the intensity of A at test is the same as on A→X trials. By the same token, when X is tested after trace conditioning, the relevant $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ and $V_{COMB\ X\rightleftarrows US}$ would be modulated by $\alpha_{A\text{-}R}S\alpha_A$ and $\alpha_{X\text{-}R}S\alpha_X$, respectively. In general, for any stimulus "S", the similarity function would be $\alpha_{S\text{-}R}S\alpha_{S\text{-}C} = (\alpha_{S\text{-}R}/(\alpha_{S\text{-}R}+|\alpha_{S\text{-}C}-\alpha_{S\text{-}R}|)) \times (\alpha_{S\text{-}C}/(\alpha_{S\text{-}C}+|\alpha_{S\text{-}C}-\alpha_{S\text{-}R}|))$; where $\alpha_{S\text{-}R} = \alpha_S$ during the test with S, and $\alpha_{S\text{-}C} = \alpha_S$ during conditioning with S.

The continuous lines in each panel show the output values for $\alpha_{X\text{-}R}S\alpha_X$ when the intensity of X during A→X trials (which determines $V_{A\rightarrow X}$ and $\alpha_{X\text{-}R}$) was the same (e.g., .50) as during X→US trials (which determines α_X), as is the case in normal higher-order conditioning procedures. Inspection of the continuous lines in panels 5a-5d confirms that the rate at which 1 is approached, across the A→X trials, decreases as α_A is reduced from .50 (panels 5a and 5b), to .30 (panel 5c), and finally to .10 (panel 5d). Inspection of the dashed lines in panels 5b-d confirm that there is a more or less extended period of A→X training trials where reducing α_X during X→US trials (from .50 to .45) results in $\alpha_{X\text{-}R}S\alpha_X$ being higher than when α_X is the same (i.e., .50; the continuous lines). This difference reverses as $\alpha_{X\text{-}R}$ (i.e., $V_{A\rightarrow X}$) approaches .50 and begins to deviate from the lower value of α_X (i.e., .45; see panels b and c, but not panel d after 10 A→X trials).

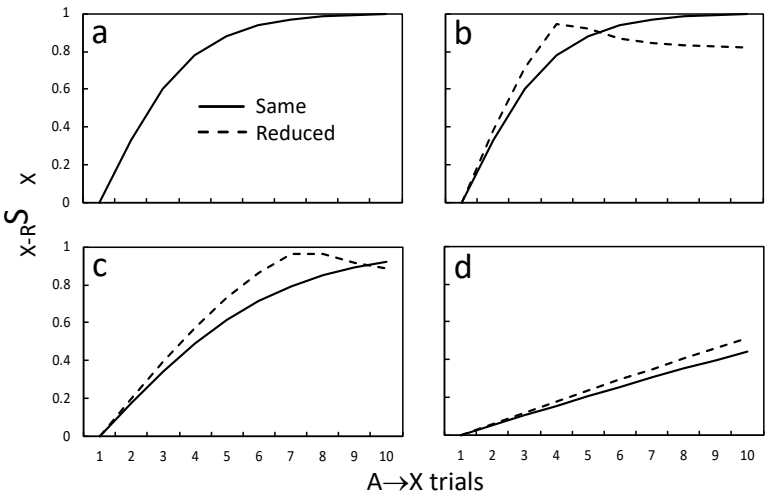

Figure 5. Computer simulations of how the number of A→X trials affects the similarity of $\alpha_{X\text{-}R}$ to α_X (i.e., $\alpha_{X\text{-}R}S\alpha_X$). The continuous lines in each panel denote output values for $\alpha_{X\text{-}R}S\alpha_X$ when the α_X value (.50) used to generate $V_{A\text{-}X}$ (i.e., $\alpha_{X\text{-}R}$) was the *same* as that for α_X on X→US trials; with α_A fixed at .50 in panels a and b, .30 in panel c, and .10 in panel d. Dashed lines denote the same output values when the α_X value used to generate $V_{A\rightarrow X}$ (i.e., $\alpha_{X\text{-}R}$; .50) was *reduced* to .45 during X→US trials. This manipulation mimics the use of trace conditioning for X→US trials.

These simulations confirm that trace conditioning (simply implemented as a reduction in stimulus intensity when X is paired with the US) has the potential to enhance

higher-order conditioning when the presentation of A results in the retrieved intensity of X (given by the numerical value of $V_{A\text{-}X}$) closely matching the conditioned intensity of X. In fact, the impact of increased similarity – through reducing α_X – on higher-order conditioning will depend on it (more than) compensating for reducing the value of $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ within the bracketed term (i.e., $V_{CHAIN\ A\rightarrow X\rightleftarrows US} + V_{COMB\ A\rightleftarrows US}$). However, reducing α_X has a relatively small effect on the rate at which X approaches the asymptote, which depends on β_{US}. Computer simulations reveal that reducing the value of α_X between the A→X trials and X→US trials by only 10% can increase R_A, R_X, and R_{US} output values by between 5% and 20% using the modified Equations 5-7. The effect of this reduction is apparent in both sensory preconditioning and second-order conditioning. Figure 6 presents a specific instance of the application of the model to sensory preconditioning, where A→X trials are followed by X→US trials before a test with A.

The upper panel of Figure 6 shows an example of how the number of A→X trials influences $\alpha_{X\text{-}R}S\alpha_X$ as a function of whether α_X is set to the same value on A→X trials and X→US trials (.40; continuous line) or is reduced by 10% from the A→X trials to the X→US (dashed line); with α_A set to .30 and β_{US} set to .50. The functions are similar to those in Figure 5. The lower panel depicts the effect of probing for sensory preconditioning after different number of A→X training trials with or without a reduction in α_X on X→US trials. In particular, it shows how the (similarity-modulated) bracketed terms within Equations 5-7 are distributed into R_A, R_X and R_{US} as a function of the number of initial A→X training trials. Inspection of the panel reveals that the output values for R_A, R_X and R_{US} for the reduced condition are initially higher than in the same condition, but this pattern reverses

in line with the reversal in the similarity function shown in the panel above.[5] R_X values are higher than the R_A values, because $V_{A\rightarrow X}$ (which also determines α_X in the performance Equations 5-7) approaches its asymptote of .40 (well within 10 A→X trials) and α_A is .30. Similarly, both R_A and R_X values are higher than R_{US} values, because β_{US} in the performance equations is given by the product of $V_{A\rightarrow X}$ and $V_{X\rightarrow US}$; even if both associations had reached asymptote the product would be less than .30 (i.e., ≈ .40 × .50).

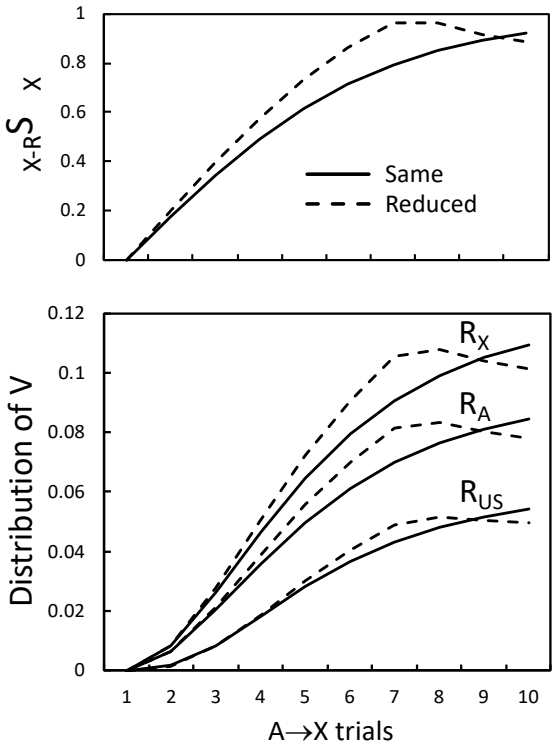

Figure 6. Computer simulations of sensory preconditioning where A is tested after different numbers of A→X trials, and the α_x value on X→US trials is either the same or reduced relative to A→X trials. The continuous line in the upper panel denotes output values for $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ when α_x was the *same* on A→X and X→US (.40) trials, and the dashed line denotes the same output values when the α_X was *reduced* from .40 on A→X trials to .35 for X→US trials. The continuous lines in the lower panel denote output values for R_A, R_X and R_{US} for the same condition, and the dashed lines denote the output values for the reduced condition. The remaining parameters were: α_A = .30 and β_{US} = .50.

Summary and integration

[5]Note that the reason that the point of reversal is actually slightly earlier for R_A, R_X, R_{US} than for similarity is because the lower value for α_X also results in slightly lower values for $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$.

Our analysis assumes that the perceived intensity of a CS (i.e., aligned to its α value) on a conditioning trial is encoded as part of the memory of the CS, and that perceived intensity provides one dimension along which generalization can occur between that CS and the same CS presented at a different intensity. Additionally, it assumes an equivalence between changing the values of α (and β) through physically changing the intensity of the stimulus (e.g., Inman et al., 2016) and changes that are dynamically generated through the processes of decay and associative retrieval (e.g., Lin & Honey, 2010; see also Iliescu et al., 2020). Thus, our analysis also provides the basis for an explanation of discrimination learning involving stimuli that differ in intensity and of timing, given the assumption that the perceived intensity of a stimulus decays lawfully across its presentation and upon its offset (cf. Iliescu et al., 2020; Staddon, 2005; Staddon & Higa, 1999). A simple modification to the base learning rules is required to implement this: Scaling the contribution of the associative strengths of stimuli to $\Sigma V_{S\text{-}TOTAL\text{-}2}$ (including $V_{1\text{-}2}$) by their similarities (subscript s) to their intensities during prior conditioning trials. For example, Equation 3 can be re-written as Equation 9, with the same similarity function as before, but within $\alpha_{X\text{-}R}S\alpha_X$, $\alpha_{X\text{-}R}$ denotes the perceived intensity of X on previous trials, while α_X is the value of the same nominal CS on the current conditioning trial. This modification captures the idea that the perceived intensity of a CS is encoded on a given conditioning trial (if α_X changes then new learning occurs), with this learning being enabled by orderly generalization of associative strength between a CS conditioned at one intensity and presented for conditioning at another (i.e., $\Sigma V_{S\text{-}TOTAL\text{-}US}$ is reduced because $\alpha_{X\text{-}R}S\alpha_X < 1$). The consequence of changing α_X from one trial to the next on the reciprocal US$\rightarrow$X association is that $V_{US\rightarrow X}$ will hone in on the new α_X – in the same way that changes in the intensity of the US over trials will change the asymptote of the X$\rightarrow$US association (see Equation 4).

$$9.\quad \Delta V_{X\rightarrow US} = \alpha x(c.\beta_{US} - \sum V_{S\text{-}TOTAL\text{-}US})$$

Finally, according to Equation 8 the similarity of different intensities of the same stimulus to one another will be symmetrical: an intense tone is as similar to a less intense tone, as the less intense tone is to the intense tone. This property of Equations 8 carries with it the implication that a discrimination between two stimulus intensities should proceed equally readily whether it is the more intense or the less intense stimulus that is followed by the US. In fact, the available evidence suggests that discrimination learning proceeds more rapidly when the more intense stimulus is paired with the US and the less intense is not, than when the roles of the two intensities are reversed (e.g., Inman et al., 2016; for a review, see Inman & Pearce, 2018). It is worth remembering, however, that according to our performance rules a more intense stimulus will elicit a greater proportion of (CS-oriented) responding than a less intense stimulus (see Equations 5-7). This characteristic of our rules has the potential to explain the asymmetry in the formation of an intensity discrimination to the extent that the measure of discrimination is more affected by CS-oriented responding than by US-oriented responding. Moreover, the value of the CS-US assembly ($V_{COMB\ CS\rightleftarrows US}$) is affected by the perceived intensities of both the CS (α_{CS}) and the US (β_{US}; i.e., $V_{COMB\ CS\rightleftarrows US} = V_{CS\rightarrow US} + (\frac{1}{c}.V_{CS\rightarrow US} \times V_{US\rightarrow CS})$), which will mean that an intense CS will result in a higher asymptote than will a CS that is less intense.

Higher-order conditioning and conditioned inhibition

One important issue deserves final consideration. Earlier we noted the conspicuous similarity between second-order conditioning and conditioned inhibition procedures. Both procedures involve reinforced X trials and nonreinforced AX trials, but they yield opposite outcomes: In second-order conditioning procedures, A acquires the capacity to generate conditioned responding, but in conditioned inhibition procedures

(where the two trials types are usually intermixed) A acquires the capacity to inhibit conditioned responding (e.g., to X). Clearly, any comprehensive integration of Pavlovian and higher-order conditioning needs to be able to generate both outcomes in a principled way. We now apply the model (including the similarity-based modulation of associative chains) to conditioned inhibition training with α_A and α_X set to a low value (.40) or the higher value (.80) that we have already noted supports second-order conditioning in the standard design (see Figure 4).

The computer simulations involved alternating reinforced X trials with nonreinforced AX trials. Of central interest was how A could possess positive values when presented alone (indicative of excitatory second-order conditioning), and the capacity to reduce the positive output values generated by X (indicative of A possessing inhibitory properties; Holland & Rescorla, 1975; Rescorla, 1976). We first need to consider not only how the the (direct) associations between A, X and the US change over the course of training, but also how the associative chains involving the same stimuli develop. The rules for calculating changes in the individual links between A and the US and X and the US follow the examples in Equations 1-4; and the formula for calculating the associative strength of the compound AX is: $V_{COMB\ AX \rightleftarrows US} = \Sigma V_{AX \rightarrow US} + (\frac{1}{c}.\Sigma_{AX \rightarrow US} \times (V_{US \rightarrow X} + V_{US \rightarrow A}))$. This formula was used in the HeiDI model (Honey et al., 2020), and follows how $V_{COMB\ A \rightleftarrows US}$ and $V_{COMB\ X \rightleftarrows US}$ are calculated for individual stimuli. We have already specified how the associative chains are calculated (e.g., $V_{CHAIN\ A \rightarrow X \rightleftarrows US} = \frac{1}{c}.V_{A \rightarrow X} \times V_{COMB\ X \rightleftarrows US}$). Our simulations make the simplifying assumption that when AX is presented, stimulus A does not activate X and X does not activate A (both are already present), and the associative chains do not contribute to performance on these AX trials (cf. Lin et al., 2013; Ward-Robinson et al., 2001); but these assumptions could be relaxed without affecting the patterns of results. We expected that $V_{COMB\ AX \rightleftarrows US}$ would tend to 0

over the course of training, and that while $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ would acquire a positive value (simulating second-order excitatory conditioning), $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ would acquire a negative value (simulating second-order conditioned inhibition; Rescorla, 1976). Figure 7a and 7b shows the results of the computer simulations after the 1st, 4th and 8th reinforced X trials, when α_A and α_X were either set to .40 (7a) or .80 (7b).

The development of direct associations and associative chains. Inspection of panel 7a confirms that following the first conditioning trial with X, $V_{COMB\ AX\rightleftarrows US}$ had a value which simply reflected the change in the output values for X alone. Similarly, the values of $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ and $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ were 0, because $V_{A\rightarrow X}$ and $V_{X\rightarrow A}$ = 0 prior to the first AX trial. However, after 4 and 8 conditioning trials (and the intervening AX trials), there was a reduction in the values of $V_{COMB\ AX\rightleftarrows US}$ (closed black squares), which reflected the impact of the inhibition acquired by A on the nonreinforced AX trials. Also, while the values of $V_{COMB\ X\rightleftarrows US}$ and $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ became increasingly positive, those of $V_{COMB\ A\rightleftarrows US}$ and $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ became increasingly negative.

Panel 7b shows that $V_{COMB\ X\rightleftarrows US}$ and $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ rapidly developed marked positive values, while the values for $V_{COMB\ A\rightleftarrows US}$ and the $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ became (mildly) inhibitory, and $V_{COMB\ AX\rightleftarrows US}$ tended to 0. $V_{COMB\ A\rightleftarrows US}$ and $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ take smaller inhibitory values in panel 7b than in panel 7a because $V_{COMB\ A\rightleftarrows US}$ is calculated by adding $V_{A\text{-}US}$ (which has a negative value) to the product of $V_{A\rightarrow US}$ and $V_{US\rightarrow A}$. This product becomes increasingly positive as $V_{US\rightarrow A}$ takes increasingly negative values, which is more likely when α_A is set to .80 than .40. The most interesting results from the simulations are shown in panels 7c and 7d, which depict the distribution of $V_{COMB\ AX\rightleftarrows US}$ into R_{AX} and R_{US} (reflecting testing with AX), of $V_{COMB\ A\rightleftarrows US}$ plus $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ into R_A, R_X and R_{US} (reflecting

testing with A), and $V_{COMB\ X\rightleftarrows US}$ plus $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ into R_A, R_X and R_{US} (reflecting testing with X).

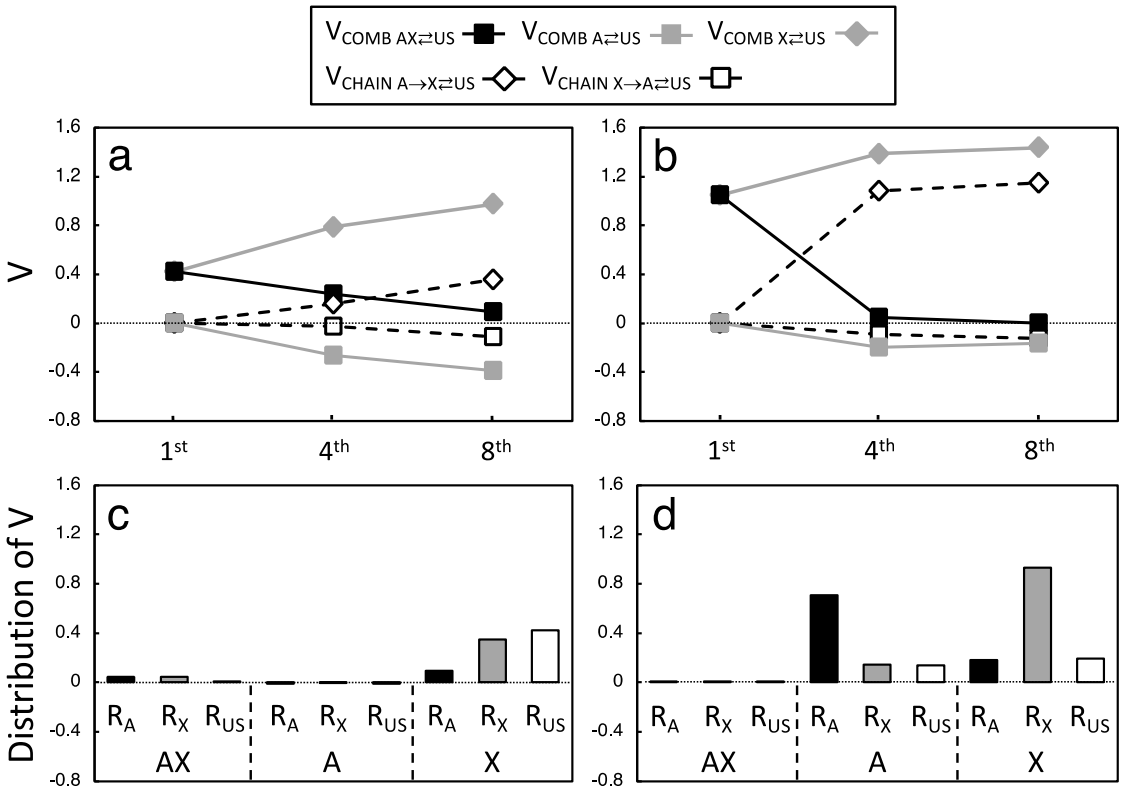

Figure 7. Computer simulations of conditioned inhibition training with alternating X→US trials and nonreinforced AX trials. α_A and α_X were set to .40 and β_{US} was set to .80 (panels a and c) or α_A, α_X and β_{US} were all set to .80 (panels b and d). Panels a and b show the output values for $V_{COMB\ AX\rightleftarrows US}$, $V_{COMB\ A\rightleftarrows US}$, $V_{COMB\ X\rightleftarrows US}$, $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ and $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ after the 1st, 4th and 8th X trial. Panels c and d show how the distribution of $V_{COMB\ AX\rightleftarrows US}$ into R_A, R_X and R_{US} for AX (from $V_{COMB\ AX\rightleftarrows US}$); from $V_{COMB\ A\rightleftarrows US}$ + $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ into R_A, R_X and R_{US} for A; and from $V_{COMB\ X\rightleftarrows US}$ + $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ into R_A, R_X and R_{US} for X. These values for the lower panels were taken after the 8th reinforced X trial.

Distribution of V after the final X conditioning trial. Panels 7c (where α_A and α_X = .40) and 7d (where α_A and α_X = .80) show the distribution of V (i.e., the bracketed term) when AX, A and X were probed after the 8th reinforced X trial. When AX was tested, $V_{COMB\ AX\rightleftarrows US}$ was distributed into R_A, R_X and R_{US} using analogues of Equations 5-7: In the proportion terms, the values of α_A and α_X were the numerators for R_A and R_X, whereas $|\frac{1}{c}.\Sigma V_{AX\rightarrow US}|$ was the numerator for R_{US}; and these values were summed for the common denominator term. When A was tested, Equations 5-7 were used to distribute the

summed value of the (inhibitory) $V_{COMB\ A\rightleftarrows US}$ and similarity-modulated (excitatory) $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$ into R_A, R_X, and R_{US}. When X was presented, the same equations were used to distribute the summed value of (excitatory) $V_{COMB\ X\rightleftarrows US}$ and the similarity-modulated (inhibitory) $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$ into R_A, R_X, and R_{US}.

The to-be-distributed Vs for panel 7c were .09 for AX ($V_{COMB\ AX\rightleftarrows US}$), -.03 for A ($V_{COMB\ A\rightleftarrows US}$ + $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$) and .86 for X ($V_{COMB\ X\rightleftarrows US}$ + $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$). Thus, while the to-be-distributed V for AX was much lower than for X, the corresponding value for A was somewhat inhibitory. Inspection of panel c shows that the values of R_{AX} and R_{US} are positive for AX, and both lower than R_X, R_A and R_{US} for X. The values for R_A, R_X and R_{US} for A were negative. The to-be-distributed Vs for panel 7d were .003 for AX ($V_{COMB\ AX\rightleftarrows US}$), .99 for A ($V_{COMB\ A\rightleftarrows US}$ + $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$) and 1.31 for X ($V_{COMB\ X\rightleftarrows US}$ + $V_{CHAIN\ X\rightarrow A\rightleftarrows US}$). Importantly, while A and X had positive Vs, the V for AX was below that for both A and X (cf. Holland & Rescorla, 1975). This is because the sum of $V_{A\rightarrow US}$ and $V_{X\rightarrow US}$ is close to zero, and so $V_{COMB\ AX\rightleftarrows US}$ is also close to zero; remembering that $V_{COMB\ AX\rightleftarrows US} = \Sigma V_{AX\rightarrow US} + (\frac{1}{c}.\Sigma_{AX\rightarrow US} \times (V_{US\rightarrow X} + V_{US\rightarrow A}))$. Similarly, while R_A, R_X and R_{US} are close to zero for AX, these values are all positive when A and X were probed alone; with R_A having the highest value when A was probed, and R_X having the highest value when X was probed.

Summary: The final piece of the theoretical puzzle has been to provide computer simulations of a procedure that is closely aligned to higher-order conditioning, but results in a quite different outcome: conditioned inhibition. We conducted these simulations in order to communicate a simple, yet important message: During conditioned inhibition training, A can acquire net inhibitory properties or net excitatory properties depending on the parameter values for α_A and α_X. Moreover, even when the parameters result A acquiring net excitatory properties (generated through an associative chain: $V_{CHAIN\ A\rightarrow X\rightleftarrows US}$), A can also reduce the excitatory properties of X when the direct (excitatory and

inhibitory) properties of the two are combined (see Figure 7). The model that we have described here, therefore, represents a promising and principled analysis of the co-existence excitatory and inhibitory processes across higher-order conditioning and conditioned inhibition training. The predictions that follow from the model are clearly testable, requiring the manipulation of CS intensity, and are independent of the measure of performance that is used (CS-oriented or US-oriented behaviors).

General Discussion

Higher-order conditioning extends the range of conditions under which the effects of Pavlovian conditioning are evident, and thereby increases its real-world significance. So, understanding the processes that underpin higher-order conditioning has translational value. However, higher-order conditioning has posed a set of fundamental challenges that has been resistant to both informal and formal theoretical treatment. The theoretical challenges include reconciling demonstrations of higher-order conditioning with those of conditioned inhibition, and providing an analysis of the conditions that affect the development of higher-order conditioning, and how it is evident in performance. Our model is built on an associative structure in which all acquired links have the potential to be reciprocal (cf. Asratian, 1965; Honey et al., 2020a). As we have already noted, such reciprocal associations provide a process by which conditioned responding (and higher-order conditioning) to a CS (A) can reflect the properties of A, and those of retrieved stimuli (e.g., X and the US; e.g., Holland, 1977; Patitucci et al., 2016; Timberlake & Grant, 1973). But they also provide a mechanism by which the presentation of the US can affect the development of associations between one CS (A) and another (X; cf. Holland, 1980a; Urcelay & Miller, 2009). Our model also includes a similarity function ($_RS_C$) that captures the relationships between the perceived intensities of a given CS, its trace and retrieved forms. This function has a number of general applications (e.g., discrimination learning;

cf. Inman et al., 2016; Lin & Honey, 2010, 2018); timing; cf. Staddon, 2005; Staddon & Higa, 1999); but for the present purposes it enables the model to provide an analysis of the conditions that generate higher-order conditioning that were not addressed by informal accounts of higher-order conditioning (e.g., Lin & Honey, 2011, 2016; Lin et al., 2013; Stout et al., 2004).

The resulting (formal) analysis of higher-order conditioning might seem complex, but it is noteworthy that it has only two free parameters in Equations 1-9 (α and β; alongside requisite decay functions), which are aligned to the perceived intensities of the stimuli. The fact that these parameters are aligned to the *perceived intensities* of stimuli allows the model to provide a potential account for both experimental manipulations (e.g., of stimulus intensity) but also individual differences in the perceived intensity of the experimenter's stimuli: α and β influence learning rates and asymptotes (Equations 1-4), and how learning (associative strength) is distributed into different behaviors that reflect the nature of the stimuli involved (Equations 5-8). The model thus explains group-level phenomena while affording the potential to account for individual differences in acquired behavior, whether that behavior originates from Pavlovian conditioning or higher-order conditioning (cf. Honey et al., 2020c): Individual differences in the vigor and form of conditioned behavior are assumed to reflect individual differences in α and β. In this context, it is worth remembering that while Pavlov reported that "*a reflex of the second order*" could be quite modest in terms of his customary measure of salivary conditioning (i.e., drops of saliva), he also noted that there were marked individual differences in the size of the effect across different experimental animals (Pavlov, 1927, p. 105). To fully assess the predictions that follow from the present model, one would need independent assays of α for A and X, and β for the US for individual animals.

Our model has clear implications for the potential use of higher-order conditioning across different domains. For example, within behavioral neuroscience, group-level differences in higher-order conditioning should be interpreted with caution: Often a single measure of conditioning is taken, and changes in this measure (contingent on some manipulation) might reflect differences in learning, or it might reflect changes in performance resulting from differences in: α (for A and X), or β (for the US) or their associated decay functions (cf. Honey & Good, 2000); or indeed the requisite computations involving the processes that these parameters represent. Returning momentarily to the results reported by Maes et al. (2020), the neurobiological manipulation might well affect the error-correction process (on AC$\rightarrow$X trials) and thereby enable the formation of a C$\rightarrow$X association; but it might also affect the way in which learning is expressed (cf. Iordanova et al., 2011).

Taking our analysis beyond behavioral and neurobiological research, the use of principles of higher-order conditioning in the clinical domain (cf. Field, 2006; Wessa & Flor, 2007) comes with health warnings: Interventions aimed at disrupting (or making use of) the processes of higher-order conditioning need to consider whether those manipulations have appropriately targeted the associative structures that underpin behavior (e.g., the associatively retrieved representation as opposed to its directly activated counterpart). They also need to consider the implications of these interventions for the various components of the associative chains, and the different behaviors that those components generate.

To conclude: Some of the underpinning ideas for our new analysis of higher-order conditioning have been formally expressed in HeiDI (Honey et al., 2020a), a model of Pavlovian learning and performance. Our new analysis identifies the requisite learning and performance rules for higher-order conditioning, and supplements those rules with a

simple function for determining the similarity between a directly activated stimulus, a decaying trace and its retrieved form (cf. Lin & Honey, 2011). The basic model is simple: 1. The perceived intensities of the stimuli that are (directly or associatively) activated at test determine how learning embodied in an extended associative network is distributed to affect performance; and 2. This process is modulated by the similarity of the perceived intensities of the test stimuli to the corresponding training stimuli within the assembly. The similarity function enables important features of higher-order conditioning to be explained, and its dynamic property affords significant additional explanatory potential, while providing a clear impetus for further experimental analysis. The resulting model provides an integrated analysis of higher-order conditioning and Pavlovian conditioning.

References

Amiro, T.W., & Bitterman, M.E. (1980). Second-order appetitive conditioning in goldfish. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *6*, 41-48.

Arcediano, F., Escobar, M., & Miller, R.R. (2005). Bidirectional associations in humans and rats. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *31*, 301-318.

Archer, T., & Sjödén, P. (1982). Higher-order conditioning and sensory preconditioning of a taste aversion with an exteroreceptive CS1. *Quarterly Journal of Experimental Psychology*, *34B*, 1-17.

Asch, S., & Ebenholtz, S.M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society, 106,* 135-163.

Asratian, E.A. (1965). *Compensatory adaptations, reflex activity, & the brain*. Oxford: Pergamon Press.

Barnet, R.C., Grahame, N.J. & Miller, R.R. (1991). Comparing the magnitudes of second-order conditioning and sensory preconditioning effects. *Bulletin of the Psychonomic Society, 29*, 133-135.

Barnet, R.C., & Miller, R.R. (1996). Second-order excitation mediated by a backward conditioned inhibitor. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *22*, 279-296.

Brogden, W.J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology, 25*, 323-332.

Cheatle, M.D., & Rudy, J.W. (1978). Analysis of second-order odor-aversion conditioning in neonatal rats: Implications for Kamin's blocking effect. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *4*, 237-249.

Cohen-Hatton, S.R., Haddon, J.E., George, D.N., & Honey, R.C. (2013). Pavlovian-to-instrumental transfer: Paradoxical effects of the Pavlovian relationship explained. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *39*, 14-23.

Cole, R.P., Barnet, R.C., & Miller, R.R. (1995). Temporal encoding in trace conditioning. *Animal Learning & Behavior, 23,* 144-153.

Cole, R.P., & Miller, R.R. (1999). Conditioned excitation and conditioned inhibition acquired through backward conditioning. *Learning and Motivation*, *30*, 129-156.

Craddock, P., Wasserman, J.S., Polack, C.W., Kosinski, T., Renaux, C., & Miller, R.R. (2018). Associative structure of second-order conditioning in humans. *Learning & Behavior*, *46*, 171-181.

Crawford, L.L., & Domjan, M. (1995). Second-order sexual conditioning in male Japanese quail (Coturnix japonica). *Animal Learning & Behavior*, *23*, 327-334.

Davey, G.C.L., & Arulampalan, T. (1982). Second-order "fear" conditioning in humans. Persistence of CR2 following extinction of CR1. *Behavior Research and Therapy, 20*, 391-396.

Davey, G.C., & Cleland, G.G. (1982). Topography of signal-centred behavior in the rat: Effects of deprivation state and reinforcer type. *Journal of the Experimental Analysis of Behavior*, *38*, 291-304.

Davey, G.C.L., & McKenna, I. (1983). The effects of post-conditioning revaluation of CS1 and UCS following Pavlovian second-order electrodermal conditioning in humans. *Quarterly Journal of Experimental Psychology, 35B*, 125-133.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology*, *49B*, 60-80.

Dickinson, A., Hall, G., & Mackintosh, N.J. (1976). Surprise and the attenuation of blocking. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*, 313-322.

Dwyer, D.M. (2012). Licking and liking: The assessment of hedonic responses in rodents. *Quarterly Journal of Experimental Psychology, 65*, 371-394.

Dwyer, D.M., Burgess, K.V., & Honey, R.C. (2012). Avoidance but not aversion following sensory-preconditioning with flavors: A challenge to stimulus substitution. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *38,* 359-368.

Dwyer, D.M., Mackintosh, N.J., & Boakes, R.A. (1998). Simultaneous activation of the representations of absent cues results in the formation of an excitatory association between them. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *24*, 163-171.

Ecker, Y., & Bar-Anan, Y. (2019). Sensory preconditioning of evaluation requires accurate memory of the co-occurrence between the neutral stimuli. *Journal of Experimental Social Psychology*, *85*, 103886.

Field, A.P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clinical Psychology Review*, 26, 857-875.

Flagel, S.B., Clark, J.J., Robinson, T.E., Mayo, L., Czuj, A., Willuhn, I., Akers, C.A., Clinton, S.M., Phillips, P.E.M., & Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. *Nature*, *469*, 53-57.

Fudim, O.K. (1978). Sensory preconditioning of flavors with a formalin-produced sodium need. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *3,* 276-285.

Gallistel, C.R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review*, *107***,** 289-344.

Gerolin, M., & Matute, H. (1999). Bidirectional associations. *Animal Learning & Behavior*, *27*, 42-49.

Gewirtz, J.C., & Davis, M. (2000). Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. *Learning & Memory*, *7*, 257-266.

Gilboa, A., Sekeres, M., Moskovitch, M., & Winocur, G. (2014). Higher-order conditioning is impaired by hippocampal lesions. *Current Biology*, *24*, 2202-2207.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1-51.

Grossberg, S. (1988). Nonlinear neural networks: Principles, Networks and Architectures. *Neural Networks*, *1*, 17-61.

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

Haselgrove, M., & Hogarth, L. (2011). *Clinical applications of learning theory*. Psychology Press: Hove.

Hearst, E., & Jenkins, H. (1974). Sign-tracking: The stimulus-reinforcer relation and directed action. Monograph of the Psychonomic Society, Austin.

Hebb, D.O. (1949). *The organization of behavior*. New York: Wiley & Sons.

Heth, C.D. (1976). Simultaneous and backward fear conditioning as a function of number of CS-UCS pairings. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *2*, 117-129.

Hoffeld, D.R., Kendall, S.B., Thompson, R.F. & Brogden, W. (1960). Effect of amount of preconditioning training upon the magnitude of sensory preconditioning. *Journal of Experimental Psychology*, *59*, 198–204.

Holland, P.C. (1977). Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *3*, 77-104.

Holland, P.C. (1980a). Second-order conditioning with and without the US. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *6*, 238-250.

Holland, P.C. (1980b). CS-US interval as a determinant of the form of Pavlovian appetitive conditioned responses. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *6*, 155-174.

Holland, P.C. (1981). Acquisition of a representation-mediated conditioned food aversion. *Learning and Motivation, 12*, 1-12.

Holland, P.C. (1983). Representation-mediated overshadowing and potentiation of conditioned aversions. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *9*, 1-13.

Holland, P.C. (1984). Origins of behavior in Pavlovian conditioning. *Psychology of Learning and Motivation*, *18*, 129-174.

Holland, P.C. (2016). Enhancing second-order conditioning with lesions of the basolateral amygdala. *Behavioral Neuroscience*, *130*, 176-181.

Holland, P.C., & Rescorla, R.A. (1975). Second-order conditioning with food unconditioned stimulus. *Journal of Comparative and Physiological Psychology, 88*, 459–467.

Honey, R.C., Dwyer, D.M., & Iliescu, A.F. (2020a). HeiDI: A model for Pavlovian learning and performance with reciprocal associations. *Psychological Review*, *127*, 829-852.

Honey, R.C., Dwyer, D.M., & Iliescu, A.F. (2020b). Elaboration of a model of Pavlovian learning and performance: HeiDI. *Current developments in associative theory: A tribute to Allan Wagner. Journal of Experimental Psychology*: *Animal Learning and Cognition*, *46*, 170-184.

Honey, R.C., Dwyer, D.M., & Iliescu, A.F. (2020c). Individual variation in the vigor and form of Pavlovian conditioned responses: Analysis of a model system. *Learning & Motivation, 72,* 101658.

Honey, R.C., Dwyer, D.M., & Iliescu, A.F. (2022). Associative change in Pavlovian conditioning: A re-appraisal. *Journal of Experimental Psychology: Animal Learning and Cognition* (in press).

Honey, R.C., & Good, M. (2000). Associative components of recognition memory. *Current Opinion in Neurobiology*, *10*, 200-204.

Honey, R.C., Good, M., & Manser, K.L. (1998). Negative priming in associative learning: Evidence from a serial-habituation procedure. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *24*, 229-237.

Honey, R.C., & Hall, G. (1991). Acquired equivalence and distinctiveness of cues using a sensory preconditioning procedure. *Quarterly Journal of Experimental Psychology*, *43B*, 121-135.

Honey, R.C., Watt, A., & Good, M. (1998). Hippocampal lesions disrupt an associative mismatch process. *Journal of Neuroscience*, *18*, 2226-2230.

Hull, C.L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Hull, C.L. (1949). Stimulus intensity dynamism (V) and stimulus generalization. *Psychological Review*, *56*, 67-76.

Iliescu, A.F., Dwyer, D.M., & Honey, R.C. (2020). Individual differences in the nature of conditioned behavior across a conditioned stimulus: Adaptation and application of a model. *Journal of Experimental Psychology*: *Animal Learning and Cognition*, *46*, 460-469.

Iliescu, A.F., Hall, J., Wilkinson, L., Dwyer, D.M., & Honey, R.C. (2018). The nature of phenotypic variation in Pavlovian conditioning. *Journal of Experimental Psychology*: *Animal Learning and Cognition*, *44*, 358-369.

Inman, R.A., Honey, R.C., & Pearce, J.M. (2016). Asymmetry in the discrimination of auditory intensity: Implications for theories of stimulus generalisation. In J.B.

Trobalon and V.D. Chamizo (Eds.) *Associative Learning and Cognition. Homage to Professor N.J. Mackintosh* (pp. 197-222). Barcelona: Edicions de la Universitat de Barcelona.

Inman, R.A., & Pearce, J.M. (2018). The discrimination of magnitude: A review and theoretical analysis. *Neurobiology of Learning and Memory*, *153*, 118-130.

Iordanova, M.D., Good, M., & Honey, R.C. (2011). Retrieval-mediated learning involving episodes requires synaptic plasticity in the hippocampus. *Journal of Neuroscience*, *31*, 7156-7162.

James, W. (1890). *The principles of psychology*. New York: Holt.

Jenkins, H., & Moore, B.R. (1973). The form of the autoshaped response with food or water reinforcer. *Journal of the Experimental Analysis of Behavior*, *20*, 163-181.

Kamil, A.C. (1969). Some parameters of the second-order conditioning of fear in rats. *Journal of Comparative and Physiological Psychology, 67*, 364-369.

Kamin, L. (1965). Temporal and intensity characteristics of the conditioned stimulus. In W.K. Prokasy (Ed.), *Classical conditioning* (pp. 118-147). New York: Appleton-Century-Crofts.

Kamin, L. (1969). Selective association and conditioning. In N.J. Mackintosh & W.K. Honig (Eds.), *Fundamental issues in associative learning* (pp. 42-89). Halifax: Dalhousie University Press.

Konorski, J. (1948). *Conditioned reflexes and neuron organization*. Cambridge: Cambridge University Press.

Lay, B.P.P., Westbrook, R.F., Glanzman, D.L., & Holmes, N.N. (2018). Commonalities and differences in the substrates underlying consolidation of first- and second-order conditioned fear. *Journal of Neuroscience*, *38*, 1926-1941.

Leyland, C.M. (1977). Higher-order autoshaping. *Quarterly Journal of Experimental Psychology*, *29*, 607-619.

Lin, T.E., Dumigan, N.M., Good, M.A., & Honey, R.C. (2016). Novel sensory preconditioning procedures identify a specific role for the hippocampus in pattern completion. *Neurobiology of Learning and Memory*, *130*, 142-148.

Lin, T.E., Dumigan, N.M., Dwyer, D.M., Good, M.A., & Honey, R.C. (2013). Assessing the encoding specificity of associations with sensory preconditioning procedures. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *39*, 67-75.

Lin, T.E., Dumigan, N.M., Recio, S.A., & Honey, R.C. (2017). Mediated configural learning in rats. *Quarterly Journal of Experimental Psychology*, *70*, 1504-1515.

Lin, T.E., & Honey, R.C. (2010). Analysis of the content of configural representations: The role of associatively evoked and trace memories. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *36,* 501-505.

Lin, T.E., & Honey, R.C. (2011). Encoding specific associative memory: Evidence from behavioral and neural manipulations. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *37*, 317-329.

Lin, T.E., & Honey, R.C. (2016). Learning about stimuli that are present and those that are not: Separable acquisition processes for direct and mediated learning. In R.A. Murphy and R.C. Honey (Eds.) *The Wiley Handbook on the Cognitive Neuroscience of Learning* (pp. 69-85). Oxford: Wiley-Blackwell.

Mackintosh, N.J. (1974). *The psychology of animal learning*. London: Academic Press.

Mackintosh, N.J. (1994). Introduction. In N.J. Mackintosh (Ed.), *The psychology of animal learning* (1-13). London: Academic Press.

Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276-298.

Mackintosh, N.J. (1976). Overshadowing and stimulus intensity. *Animal Learning & Behavior*, *4*, 186-192.

Mackintosh, N.J. (1994). Introduction. *Animal learning and cognition*. Academic Press.

Maes, E.J.P., Sharpe, M.J., Usypchuk, A., Lozzi, M., Gardner, M.P.H., Chang, C.Y., Schoenbaum, G., & Iordanova, M.D. (2020). Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors. *Nature Neuroscience*, *23*, 176-178.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

McLaren, I.P.L., Kaye, H., & Mackintosh, N.J. (1989). An associative theory of the representation of stimuli: applications to perceptual learning and latent inhibition. In R.G.M. Morris (Ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (pp. 102-130). Oxford: Clarendon Press.

Miller, R.R., & Barnet, R.C. (1993). The role of time in elementary associations. *Current Directions in Psychological Science*, *2*, 106-111.

Mollick, J.A., Hazy, T.E., Krueger, K.A., Nair, A., Mackie, P., Herd, S.E., & O'Reilly, R.C. (2020). A systems neuroscience model of phasic dopamine. *Psychological Review*, *6*, 972-1021.

Nairne, J.S., & Rescorla, R.A. (1981). Second-order conditioning with diffuse auditory reinforcers in the pigeon. *Learning and Motivation*, *12*, 65-91

Nasser, H.M., Chen, Y-W., Fiscella, K., & Calu, D.J. (2015). Individual variability in behavioral flexibility predicts sign-tracking tendency. *Frontiers in Behavioral Neuroscience*, *9*, 289.

Navarro, V.M., & Wasserman, E.A. (2020). Bidirectional conditioning: Revisiting Asratyan's 'alternating' training technique. *Neurobiology of Learning and Memory*, *171*, 107211.

Patitucci, E., Nelson, A.J.D., Dwyer, D.M., & Honey, R.C. (2016). The origins of individual differences in how learning is expressed in rats: A general-process perspective. *Journal of Experimental Psychology*: *Animal Learning and Cognition*, *42*, 313-324.

Pavlov, I.P. (1927). *Conditioned Reflexes*. London: Oxford University Press.

Pavlov, I.P. (1949). *Collected works, Vol. III*. Moscow-Leningrad.

Pearce, J.M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587-607.

Pearce, J.M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*, 73-95.

Pearce, J.M., & Hall G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychological Review*, *87*, 532-552.

Prewitt, E. P. (1967). Number of preconditioning trials in sensory preconditioning using CER training. *Journal of Comparative and Physiological Psychology, 64*, 360-362.

Rashotte, M. (1981). Second-order autoshaping: Contributions to the research and theory of Pavlovian reinforcement by conditioned stimuli. In C.M. Locurto, H.S. Terrace, and J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 139-180). New York: Academic Press.

Rashotte, M., Griffin, R.W., & Sisk, C.L. (1977). Second-order conditioning of the pigeon's keypeck. *Animal Learning & Behavior, 5*, 25-38.

Rescorla, R.A. (1969). Pavlovian conditioned inhibition. *Psychological Bulletin*, *72*, 77-94.

Rescorla, R.A. (1976). Second-order conditioning of Pavlovian conditioned inhibition. *Learning and Motivation*, *7*, 161-172.

Rescorla, R.A. (1980). Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology*: *Animal Behavior Processes, 6*, 207-216.

Rescorla, R.A. (1982). Simultaneous second-order conditioning produces S-S learning in conditioned suppression. *Journal of Experimental Psychology*: *Animal Behavior Processes, 8*, 23-32.

Rescorla, R.A. (2000). Associative changes in excitors and inhibitors differ when they are conditioned in compound. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *26*, 428-438.

Rescorla, R.A., & Cunningham, C.L. (1978). Within-compound flavor associations. *Journal of Experimental Psychology*: *Animal Behavior Processes, 4*, 267-275.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy (eds) *Classical conditioning II*: *Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rhodes, S.E.V., Creighton, G., Killcross, A.S., Good, M., & Honey, R.C. (2009). Integration of geometric with luminance information in the rat: Evidence from within-compound associations. *Journal of Experimental Psychology*: *Animal Behavior Processes*, *35*, 92-98.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by back-propagating errors. In D.E. Rumelhart & J.L.

McClelland (Eds.), *Parallel distributed processing* (Vol. 1, chapter 8). Cambridge, M.A: MIT Press.

Rizley, R.C., & Rescorla, R.A. (1972). Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology, 81*, 1-11.

Robinson, T.E., & Flagel, S.B. (2009). Dissociating the predictive and incentive motivational properties of reward-related cues through the study of individual differences. *Biological Psychiatry*, *65*, 869-873.

Scavio, M.J., & Gormezano, I. (1974). CS intensity effects on rabbit nictitating membrane conditioning, extinction and generalization. *Pavlovian Journal of Biological Science*, *9*, 25-34.

Schultz, W., Dayan, P., & Montague, R.R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1599.

Seymour, B., Doherty, J.P., Dayan, P., Koltzenburg, M., Kones, A.K., Dolan, R.J., Friston, K.J., & Frackowiak, R.S. (2004). Temporal-difference models describe higher-order learning in humans. *Nature*, *429*, 664-667.

Shanks, D.R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press.

Silva, A.I., Haddon, J.E., Trent, S., Syed, Y., Lin, T-C, E., Patel, Y., Carter, J., Haan, N., Honey, R.C., Humby, T., Assaf, Y., Linden, D.E., Owen, M.J., Ulfarsson, M.O., Stefansson, H., Hall, J., & Wilkinson, L.S. (2019). *Cyfip1* haploinsufficiency is associated with white matter changes, myelin thinning, reduction of mature oligodendrocytes and behavioural inflexibility. *Nature Communications*, *10*, 3455.

Staddon, J.E.R. (2005). Interval timing: memory, not a clock. *Trends in Cognitive Sciences, 9*, 312-314.

Staddon, J.E.R. & Higa, J.J. (1999). Time and memory: towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior, 71*, 215-251.

Stanhope, K.J. (1992). The representation of the reinforcer and the force of the pigeon's keypeck in first- and second-order conditioning. *Quarterly Journal of Experimental Psychology*, *44B*, 137-158.

Stout, S., Escobar, M. & Miller, R.R. (2004). Trial number and compound stimuli temporal relationship as joint determinants of second-order conditioning and conditioned inhibition. *Animal Learning & Behavior, 32*, 230-239.

Stout, S.C., & Miller, R.R. (2007). Sometimes-Competing Retrieval (SOCR): A Formalization of the Comparator Hypothesis. *Psychological Review*, *114*, 759-783.

Sutton, R.S., & Barto, A.G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*, 135-170.

Tait, R.W., & Saladin, M.E. (1986). Concurrent development of excitatory and inhibitory associations during backward conditioning. *Animal Learning & Behavior, 14***,** 133-137.

Timberlake, W., & Grant, D.L. (1975). Auto-Shaping in rats to the presentation of another rat predicting food. *Science*, *190*, 690-692.

Urcelay, G.P., & Miller, R.R. (2009). Potentiation and overshadowing in Pavlovian fear conditioning. *Journal of Experimental Psychology. Animal Behavior Processes*, *35*, 340-356.

Van Hamme, L.J., & Wasserman, E.A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, *25*, 127-151.

Wagner, A.R. (1981). SOP: A model of automatic memory processing in animal behavior. In N.E. Spear & R.R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5-48). Hillsdale, NJ: Erlbaum.

Wagner, A.R., Logan, F.A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology*, *76*, 171-180.

Wagner, A.R., & Rescorla, R.A. (1972). Inhibition in Pavlovian conditioning: Applications of a theory. In R.A. Boakes & M. S. Haliday (Eds.), *Inhibition and learning* (pp. 301-336). New York: Academic Press.

Ward-Robinson, J. (2004). An analysis of second-order autoshaping. *Learning and Motivation*, *35*, 1-21.

Ward-Robinson, J., Coutureau, E., Good, M., Honey, R.C., Killcross, A.S., & Oswald, C.J.P. (2001). Excitotoxic lesions of the hippocampus leaves sensory preconditioning intact: Implications for models of hippocampal function. *Behavioral Neuroscience*, *115*, 1357-1362.

Ward-Robinson, J., & Hall, G. (1996). Backward sensory preconditioning. *Journal of Experimental Psychology*: *Animal Behavior Processes, 22*, 395-404.

Ward-Robinson, J., & Hall, G. (1998). Backward sensory preconditioning when reinforcement is delayed. *Quarterly Journal of Experimental Psychology*, *51B*, 349-362.

Warren, H.C. (1921). *A history of the association psychology*. New York: Charles Scribner's Sons.

Wessa, M., & Flor, H. (2007). Failure of extinction of fear responses in post-traumatic stress disorder: Evidence from second-order conditioning. *American Journal of Psychiatry*, *164*, 1684-1692.

Wimmer, G.E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science*, *338*, 270-273.

Yu, T., Lang, S., Birbaumer, N., & Kotchoubey, B. (2014). Neural correlates of sensory preconditioning: A preliminary fMRI investigation. *Human Brain Mapping*, *35*, 1297-1304.

Zentall, T.R., Sherburne, L.M., & Steirn, J.N. (1992). Development of excitatory backward associations during the establishment of forward associations in a delayed conditional discrimination by pigeons. *Animal Learning & Behavior*, *20*, 199-206.