frontiers in Behavioral Neuroscience

REVIEW
published: 09 September 2021
doi: 10.3389/fnbeh.2021.726218

Higher-Order Conditioning: What Is Learnt and How it Is Expressed

Robert C. Honey and Dominic M. Dwyer*

School of Psychology, Cardiff University, Cardiff, United Kingdom

Pairing a neutral conditioned stimulus (CS) with a motivationally significant unconditioned stimulus (US) results in the CS coming to elicit conditioned responses (CRs). The widespread significance and translational value of Pavlovian conditioning are increased by the fact that pairing two neutral CSs (A and X) enables conditioning with X to affect behavior to A. There are two traditional informal accounts of such higher-order conditioning, which build on more formal associative analyses of Pavlovian conditioning. But, higher-order conditioning and Pavlovian conditioning have characteristics that are beyond these accounts: Notably, the two are influenced in different ways by the same experimental manipulations, and both generate conditioned responses that do not reflect the US *per se*. Here, we present a formal analysis that sought to address these characteristics.

Keywords: association, behavior, Pavlovian conditioning, similarity, timing

OPEN ACCESS

Edited by:
Nathan Holmes,
University of New South Wales, Australia

Reviewed by:
Youcef Bouchekioua,
Hokkaido University, Japan
R. Fred Westbrook,
University of New South Wales, Australia

****Correspondence:***
Robert C. Honey
honey@cardiff.ac.uk

Specialty section:
This article was submitted to Learning and Memory, a section of the journal Frontiers in Behavioral Neuroscience

Received: *16 June 2021*
Accepted: *19 July 2021*
Published: *09 September 2021*

Citation:
Honey RC and Dwyer DM (2021) Higher-Order Conditioning: What Is Learnt and How it Is Expressed. Front. Behav. Neurosci. 15:726218. doi: 10.3389/fnbeh.2021.726218

INTRODUCTION

Pavlov observed that dogs given pairings of light with food came to salivate during the light, but also during a tone that was later paired with the light. In his terms, the light (a conditioned stimulus, CS) had become a substitute for food (an unconditioned stimulus, US), evidenced both through the capacity of the light to elicit salivation (the conditioned response, CR) and to support a "*reflex of the second order*" to the tone. In fact, Pavlov described such second-order CRs as "*in most cases very weak*," indicating that there were substantial individual differences in their size and transience (Pavlov, 1927; pp. 104–105). We will return to the important issue of individual differences towards the end of this article. For now, it is sufficient to note that second-order conditioning is a well-established phenomenon across a range of preparations (e.g., *appetitive conditioning*: Rashotte et al., 1977; *aversive conditioning*: Rizley and Rescorla, 1972; *sexual conditioning*: Crawford and Domjan, 1995), and so too is another example of higher-order conditioning, sensory preconditioning (e.g., *appetitive conditioning*: Allman and Honey, 2006; *aversive conditioning*: Brogden, 1939; *flavor-aversion learning*: Rescorla and Cunningham, 1978). For sensory preconditioning, the tone and light in the opening example are paired before the light is conditioned, whereupon the tone also elicits conditioned responding (see **Table 1**).

Higher-order conditioning procedures have become a popular means of examining the neurobiology of learning and memory (for a review, see Gewirtz and Davis, 2000; see also, e.g., Lin and Honey, 2011; Gilboa et al., 2014; Holland, 2016; Lin et al., 2016; Lay et al., 2018; Maes et al., 2020; Mollick et al., 2020). This popularity reflects the relevance of higher-order conditioning to clinical domains (e.g., Davey and Arulampalan, 1982; Davey and McKenna, 1983; Wessa and Flor, 2007; see also, Field, 2006; Haselgrove and Hogarth, 2011), but also the practical advantages of the procedures, and the potential insights that their use enables: The procedures allow the complex

TABLE 1 | Higher-order conditioning procedures.

	Stage 1	Stage 2	Test
Second-order conditioning:	X→US	A→X	A?
Sensory preconditioning:	A→X	X→US	A?

Note: A and X are conditioned stimuli and the US denotes an unconditioned stimulus.

effects generated by the presentation of a motivationally significant US, on X→US trials, to be separated from the associative processes operating on A→X trials; and they also allow the nature of different acquisition and performance processes to be separately probed. But, what is learned during higher-order conditioning and how is that learning expressed? These two related questions have not been addressed in an integrated fashion by traditional accounts of higher-order conditioning. In fact, a recent critical review of evidence relating to these accounts suggested that they leave many important issues unresolved, which motivated the development of a new computational model of higher-order conditioning (Honey and Dwyer, under review). This model was built on a recent analysis of Pavlovian conditioning and performance: HeiDI (Honey et al., 2020a). Here, we first present a synthesis of extant informal accounts of higher-order conditioning together with the evidence that they fail to address, before presenting the new computational model of higher-order conditioning.

TRADITIONAL ACCOUNTS OF HIGHER-ORDER CONDITIONING

Mackintosh (1974; pp. 85–91; see also Gewirtz and Davis, 2000) identified two accounts of higher-order conditioning that have enjoyed an enduring appeal. One is closely aligned to conventional accounts of Pavlovian conditioning, wherein an association is held to form between the CS representation and either the US representation (i.e., a stimulus-stimulus association) or the processes responsible for responses that it generates (i.e., a stimulus-response association). For higher-order conditioning, it has been argued that an association forms between stimulus A and the US (or processes involved in generating the CR) through a process of representation mediated learning. Thus, for second-order conditioning, the X→US trials might allow A to become linked to the representation of the US that is retrieved by X on A→X trials (Konorski, 1948, p. 68) or to processes more directly responsible for the CR to X (Pavlov, 1927, p. 105; Rizley and Rescorla, 1972). Whereas for sensory preconditioning, the A→X trials might allow the representation of A retrieved by X on X→US trials to be linked to the US (e.g., Ward-Robinson and Hall, 1996, Ward-Robinson and Hall, 1998; see also, Holland, 1981; Hall, 1996; Iordanova et al., 2011). Accounts based upon representation mediated learning are often contrasted with the simpler possibility that a (directional) associative chain underpins higher-order conditioning (e.g., Gewirtz and Davis, 2000). Here, X→US pairings allow an association to form between representations of X and the US, or those processes responsible for the UR, while A→X pairings enable an association to develop between representations of A and X. The efficacy of the associative chains, A→X→US or A→X→UR, will then determine the propensity for A to elicit conditioned responding. However, the accounts described above and depicted in **Figure 1** are challenged by the conditions under which higher-order conditioning is observed and how it is evident in behavior.

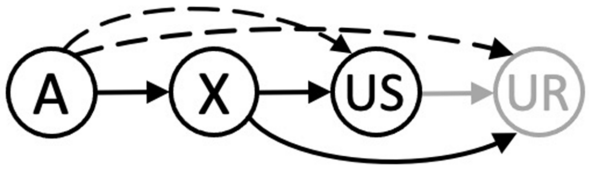

FIGURE 1 | An integrated schematic for higher-order conditioning: The associative chains (black continuous arrows) and retrieval mediated associations (black dashed arrows) resulting from separate A→X and X→US trials. The gray solid arrow is an (unconditioned) link between the US its UR. The A, X, and US nodes are held to be activated by their corresponding stimuli, with the UR generated by the presentation of the US.

A SYNTHESIS OF UNRESOLVED ISSUES

The Conditions Under Which Higher-Order Conditioning Is Observed

When there is a trace interval between a CS and US (i.e., X→trace→US), conditioned responding during the CS is normally less evident than when there is no interval (see Mackintosh, 1983, pp. 86–89). The accounts of higher-order conditioning outlined above seem constrained to predict that when there is a trace interval between X and the US the CR to A should also be less marked: X→trace→US trials will be an ineffective basis for X to retrieve the US (or evoke the UR) on A→X trials in second-order conditioning procedures, and X→trace→US will be an ineffective vehicle for the retrieved representation of A to become linked to the US in sensory preconditioning procedures. Similarly, the final X→US or X→UR link in any associative chain will be less effective (in both procedures) after X→trace→US trials. However, trace conditioning with X enhances conditioned responding to A in both sensory preconditioning (Ward-Robinson and Hall, 1998; Lin and Honey, 2011; see also, Kamil, 1969) and second-order conditioning procedures (Lin and Honey, 2011; see also, Cole et al., 1995; Barnet and Miller, 1996). Another simple observation is similarly problematic: Extinguishing first-order conditioned responding to X, before test trials with A, does not (always) reduce the capacity of A to generate responding in sensory preconditioning (Ward-Robinson and Hall, 1996) or second-order conditioning procedures (e.g., Rizley and Rescorla, 1972; Cheatle and Rudy, 1978; Amiro and Bitterman, 1980; Nairne and Rescorla, 1981; Archer and Sjödén, 1982; but see Rescorla, 1982). These results are inconsistent with an associative chain account to the extent that the efficacy of the final link in the chain should have been reduced by extinguishing X, and they have been taken to support the view that A has an association with the US (or its UR) that is independent of the association of X with the US (or its UR). A final intriguing observation about

sensory preconditioning is that when A is presented together with X during the test, the resulting AX compound provokes more conditioned responding than when X is either presented alone or with a control stimulus (e.g., Ward-Robinson et al., 2001; Lin et al., 2013). By default, and ignoring the results from the trace conditioning procedure, these results have been taken to support a retrieval mediated learning account since it supposes that A has a basis to elicit conditioning responding independently of X. However, these results could also reflect the fact that the directly activated representation of a stimulus (X), and its trace or retrieved representations (X*; see Lin and Honey, 2011, 2016; Lin et al., 2013) can be discriminated from one another, and enter into separate associations that affect performance in distinct ways (Lin and Honey, 2010). For example, enhanced higher-order conditioning with trace conditioning could reflect the fact that the representation of X that is retrieved by A is more similar to the representation of X that enters into association with the US during trace conditioning than during standard conditioning. Also, whether the extinction of X does or does not affect responding to A could be determined by the similarity of the representation of X retrieved by A during the test to the representation of X that was subject to extinction (see Rescorla, 1982). Later, we will develop a more formal analysis of this suggestion, which relies on representations of X, its trace and retrieved forms being dynamically coded in terms of the dimension of perceived intensity, and forming part of what is learned about a given stimulus.

How Higher-Order Conditioning Is Evident in Behavior

Higher-order conditioning procedures include two types of trial, A→X and X→US, and there has been an understandable focus on how X→US trials enable responding to A. However, A→X trials can—in and of themselves—generate behavior. For example, when an auditory stimulus is paired with a localized visual stimulus (i.e., A→X), A comes to elicit an orienting response that reflects the location in which X is presented (e.g., Honey et al., 1998a,b; see also, Narbutovich and Podkopayev, 1936; cited in Konorski, 1948, p. 91; Silva et al., 2019). Any complete analysis of higher-order conditioning needs to address the fact that A will come to elicit behaviors that reflect the nature of both the US and X (see Lin and Honey, 2011, 2016; Lin et al., 2013). Not considering how the nature of the retrieved X might affect behavior to A is a pervasive issue with both informal accounts of higher-order conditioning and more formal models of Pavlovian conditioning: How do the proposed associative structures generate different forms of behavior? This process has been left underspecified by both formal models of Pavlovian conditioning (e.g., Mackintosh, 1975; Rescorla and Wagner, 1972; Pearce and Hall, 1980; Wagner, 1981) and informal accounts of higher-order conditioning.

The accounts of higher-order conditioning that we have considered assume that the associations responsible for performance are directional. For accounts based on representation mediated learning, the association is from A to the US (i.e., A→US), whereas for those based on an associative chain they are from A to X (i.e., A→X) and from X to the US (i.e., X→US). The requisite additional assumption is that performance is (ordinally) related to either the strength of the association between A and the US (i.e., $V_{A\text{-}US}$), or the product of the links in the associative chain (i.e., $V_{A\text{-}X\text{-}US} = V_{A\text{-}X} \times V_{X\text{-}US}$; see Rescorla and Wagner, 1972). But, we know that accounts based on such assumptions are, at best, incomplete: The conditioned behavior generated by X→US trials reflects both the properties of the US and of the CS (e.g., Timberlake and Grant, 1975; see also, Holland, 1977; Patitucci et al., 2016; Iliescu et al., 2018). In fact, following Holland (1977, 1984), we can broadly distinguish between CS-oriented conditioned responding (e.g., sign-tracking; Hearst and Jenkins, 1974; see also, Davey and Cleland, 1982; Flagel et al., 2009) and US-oriented responding (e.g., goal-tracking; Boakes, 1977). Directional associations or chains of such associations from a CS to the US provide no foundation for CS-oriented conditioned behaviors[1]. Similarly, behaviors generated through Pavlovian conditioning (e.g., X→US) are not (quantitatively or qualitatively) the same as those generated by higher-order conditioning trials (e.g., A→X). This should be so if higher-order conditioned behavior is generated solely by associative activation of the US representation (see Holland and Rescorla, 1975; see Pavlov, 1927). Two examples from quite different preparations will suffice.

Stanhope (1992) gave hungry and thirsty pigeons training where keylight X was paired with food and keylight Y was independently paired with water. As a result, the pigeons directed pecks to X and Y, but those to X (the food keylight) were of greater force than those to Y (the water keylight; see Jenkins and Moore, 1973). The pigeons were then given trials where keylight A was paired with X while B was paired with Y. As a result, A and B came to elicit keypecking (see Rashotte et al., 1977), but the force of the keypecks to A and B did not differ in force (see also, e.g., Holland, 1977). Dwyer et al. (2012) gave thirsty rats separate access to two flavor compounds containing two flavors (A with X and B with Y); and then rats received access to X paired with illness and access to Y that was not. This procedure resulted in a reluctance to consume X relative to Y, and also A relative to B (see Rescorla and Cunningham, 1978). An important further finding was that while the first-order aversion was also evident in how rats consumed X (i.e., as a reduction in lick cluster size, indicative of a reduction in hedonic responses; see Dwyer, 2012), the second-order aversion to A was not. Neither a mediated A→US association nor an A→X→US associative chain provides a principled basis for the dissociations observed by Stanhope (1992) and by Dwyer et al. (2012; see also Holland and Rescorla, 1975).

A MORE FORMAL ANALYSIS

The model that we now describe builds on the assumption that learning involves the development of reciprocal associations: a central feature of the HeiDI model (see Honey et al., 2020a,b,c). This assumption provides a basis for the fact that conditioning

[1]Some combinations of stimuli might activate response units that generate behaviors that do not closely resemble those observed when the same stimuli are presented individually (e.g., conditioned freezing).

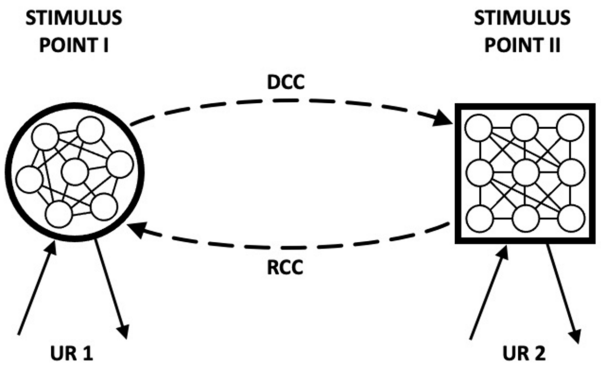

FIGURE 2 | "Figure 8 Scheme of the conditioned reflex with bidirectional connection. DCC, direct conditioned connection; RCC, reverse conditioned connection; UR 1, unconditioned reflex No. 1; UR 2, unconditioned reflex No. 2." Adapted from Asratian (1965). Note that while Stimulus Points I and II are reciprocally connected, the elements within Stimulus Points I and II are neither fully interconnected with one another and nor is their nature made explicit.

can result in both an increase in CS-oriented and US-oriented behaviors to a CS, and was foreshadowed by Asratian (1965). **Figure 2** is an adaptation of Figure 8 (Asratian, 1965; p. 179) where standard conditioning trials are held to result in a directly conditioned connection (DCC) and reverse conditioned connection (RCC) between Stimulus Points I and II (e.g., A and X, or X and the US). UR 2 can be generated both through direct activation of Stimulus Point II and through DCC by activation of Stimulus Point I, and UR 1 can be generated through activation of Stimulus Point 1 and by activation of Stimulus Point II through RCC. There is evidence to support the idea that reciprocal associations are formed during CS→US pairings (e.g., Asch and Ebenholtz, 1962; Heth, 1976; Tait and Saladin, 1986; Zentall et al., 1992; Gerolin and Matute, 1999; Arcediano et al., 2005).

The model described here and developed in Honey and Dwyer (under review), has three components: (1) Learning rules together with the associative structures that they generate; (2) performance rules that determine how those structures generate different behaviors; and (3) a function that specifies the similarity between a CS, its trace, and retrieved forms, in terms of their perceived intensities. Schematics for the associative structures generated by higher-order conditioning trials (i.e., A→X and X→A) are depicted in **Figure 3**. We assume that the unconditioned structure has existing links of differing strengths from A, X, and the US to a set of response units (r1-r6; left panel), and that reciprocal (excitatory) links form between A and X, and between X and the US during both sensory preconditioning (middle panel) and second-order conditioning (right panel). In the case of sensory preconditioning conditioning, the X→US trials will also result in the formation of an accompanying inhibitory US→A link, whereas in the case of second-order conditioning, the A→X trials result in the formation of an inhibitory A→US link (see next paragraph).

In general terms, the formation of reciprocal links between the components of higher-order conditioning trials (A, X, and the US) provides a mechanism by which conditioned responding (to X) and higher-order conditioning (to A) are affected by the properties of the components of any given trial. In the case of higher-order conditioning, performance during A will reflect its properties (e.g., Holland, 1977; Patitucci et al., 2016; Iliescu et al., 2018), and those of the stimuli with which it is associated: X (Honey et al., 1998a,b; Silva et al., 2019; see also, Narbutovich and Podkopayev, 1936; cited in Konorski, 1948, p. 91) and the US (e.g., Holland and Rescorla, 1975; Holland, 1977; Stanhope, 1992; Dwyer et al., 2012). Similarly, performance to X will reflect the stimulus itself as well as its associations with A and the US. The issue then becomes one of specifying how the combined associative strengths within the extended associative structures (see **Figure 3**) is distributed to reflect the properties of A through the response units it is connected to and those of the retrieved representations of X and US. Following HeiDI (Honey et al., 2020a), we assume that they do so in proportion to their perceived intensities: for example, if the perceived intensity of A is higher than that of the retrieved memories of X or the US then a greater proportion of the combined associative strength would generate responses that are linked to A. Finally, we assume that this process is modulated by the similarity between the perceived intensities of the stimuli presented at the test (e.g., the associatively retrieved memory of X) to their perceived intensities on the conditioning trials (see Ward-Robinson and Hall, 1996, 1998; Ward-Robinson et al., 2001; Lin and Honey, 2011, 2016; Lin et al., 2013; see also, Kamil, 1969; see also, Cole et al., 1995; Barnet and Miller, 1996). We now give formal expression to these general ideas.

Learning Rules

The formation of reciprocal associations between stimulus 1 and stimulus 2, having perceived intensities of α_1 and α_2, is determined by two equations: $\Delta V_{1\text{-}2} = \alpha_1(c.\alpha_2 - \Sigma V_{TOTAL\text{-}2})$; and $\Delta V_{2\text{-}1} = \alpha_2(c.\alpha_1 - \Sigma V_{TOTAL\text{-}1})$[2]. These rules underpin the HeiDI model (Honey et al., 2020a). For both equations, associative changes on a given trial ($\Delta V_{1\text{-}2}$ and $\Delta V_{2\text{-}1}$) are influenced by pooled error terms (i.e., $c.\alpha_2 - \Sigma V_{TOTAL\text{-}2}$ and $c.\alpha_1 - \Sigma V_{TOTAL\text{-}1}$) in which $\Sigma V_{TOTAL\text{-}2}$ and $\Sigma V_{TOTAL\text{-}1}$ are the summed associative strengths of stimuli present on that trial to the subscripted stimulus ($_1$ or $_2$). The maximum possible associative strengths are given by c (which is 1 in units of V) multiplied by the perceived intensities of the stimuli (α_2 and α_1)[3]. Otherwise, the learning rules are simplified extensions to the one developed by Rescorla and Wagner (1972; see also, McLaren et al., 1989)[4]. Equations 1 and 2 reference these generic equations to the critical A→X and X→A associations, and Equations 3 and 4 reference them to the X→US and

[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).

[3]The fact that the asymptotes and the rates at which they are reached are determined by α_1 and α_2 creates computational advantages when specifying the similarity of (1) the retrieved values of α_2 and α_1 (given by the numerical values of $V_{1\text{-}2}$ and $V_{2\text{-}1}$, respectively), and (2) their conditioned values α_2 and α_1.

[4]The rules have no independent lambda (λ) parameter to determine the asymptote for the $V_{1\text{-}2}$ association (or for the $V_{2\text{-}1}$ association). There is also no need to have separate learning rate parameters for when the target for the association (1 or 2) is present (e.g., β_E) and absent (e.g., β_I; see Honey et al., 2020b). β_I was required by the Rescorla-Wagner model $-\Delta V_{CS\text{-}US} = \alpha\beta(\lambda - \Sigma V)$ — to enable learning to occur when the US was absent and β would otherwise = 0.

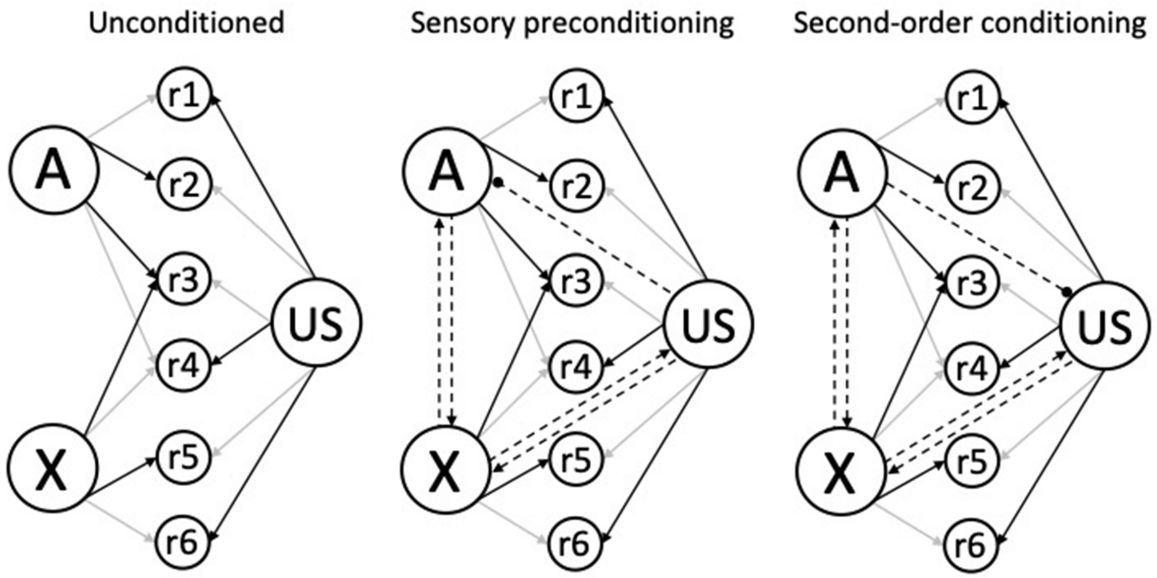

FIGURE 3 | Schematic associative structures for higher-order (excitatory) conditioning. The left structure shows the unconditioned links from the CSs (A and X) and the US to response-generating units (r1-r6) before conditioning. The darkness of the arrows indicates link strength: A is strongly linked to r2 and r3, B is strongly linked to r3 and r5; and the US is strongly linked to r1, r4, and r6; and the remaining unconditioned links are weak or absent. The central and right structures show the reciprocal associations between the A and X, and between the X and US nodes (denoted by the dashed lines with arrowheads), acquired during higher-order trials (e.g., A→X and X→US); with a directional inhibitory US→A association for sensory preconditioning (center panel) and an inhibitory A→US association in second-order conditioning (right panel; denoted by the dashed line with the circular end; based upon one interpretation of inhibitory learning). Adapted from Honey and Dwyer (under review).

US→X associations (analogous equations can be specified for the reciprocal links between A and the US). The maximum associative strength in Equation 3 is set by $\beta_{US,}$ which is the learning rate parameter in Equation 4.

$$\Delta V_{A\text{-}X} = \alpha_A(c.\alpha_X - \Sigma V_{TOTAL\text{-}X}) \quad (1)$$

$$\Delta V_{X\text{-}A} = \alpha_X(c.\alpha_A - \Sigma V_{TOTAL\text{-}A}) \quad (2)$$

$$\Delta V_{X\text{-}US} = \alpha_X(c.\beta_{US} - \Sigma V_{TOTAL\text{-}US}) \quad (3)$$

$$\Delta V_{US\text{-}X} = \beta_{US}(c.\alpha_X - \Sigma V_{TOTAL\text{-}X}) \quad (4)$$

This analysis already affords additional explanatory power in the context of demonstrations of higher-order conditioning. For example, the analysis provides a simple explanation for (so-called) backward sensory preconditioning (Ward-Robinson and Hall, 1996, 1998). In this case, the fact that X→A pairings replace the typical A→X pairings has been taken to mean that an A→X→US chain cannot be constructed upon which to generate conditioned responding to A. The suggestion that X→A pairings enable reciprocal associations to form between X and A means that an A→X→US associative chain is generated. The same form of argument can be applied to the fact that when the usual X→US trials are replaced with US→X trials, subsequent presentations of A provoke marked (US-oriented) responding in a sensory preconditioning procedure (for an alternative analysis, see Miller and Barnet, 1993; see also, Cole and Miller, 1999). Finally, it has been demonstrated that second-order conditioning to A is reduced if the US is presented on the A→X trials (i.e., A→X→US; see Holland, 1980). This result is predicted to the extent that the US competes with A to become associated with X (because it is more intense; Mackintosh, 1976) and with X to become associated with A; and that this reduction in the strength of the A→X association outweighs the fact that X continues to be paired with the US.

Performance Rules

Having specified the learning rules that generate the associative structures depicted in **Figure 3**, we now need to specify how these structures give rise to different conditioned behaviors. Our analysis is again based on HeiDI (Honey et al., 2020a,b,c). HeiDI separates the associative strengths of the CS→US and US→CS associations (Hebb, 1949) from the influence on performance of the intensities of the (presented) CS and (retrieved) US (see Hull, 1949). Thus, when the CS is presented the combined strength of the reciprocal associations [$V_{COMB} = V_{CS\text{-}US}$ + (numerical value of $V_{CS\text{-}US} \times V_{US\text{-}CS}$)] is distributed into CS- and US-oriented components (R_{CS} and R_{US}, respectively).[5] With this distribution being determined by the perceived intensity of the CS (α_{CS}) relative to the (retrieved) US (β_{US}, as retrieved by the CS; see Holland, 1977; Patitucci et al., 2016). In general, this means that when α_{CS} is higher than β_{US}, the CS-oriented component

[5]The reciprocal associations are combined in this way, rather than being simply mapped onto CS-oriented (US→CS) and US-oriented (CS→US) responding, to reflect the interactive nature of the reciprocal associations, but also to avoid the prediction that extinction of the CS would leave CS-oriented responding unaffected because it would only impact the CS→US association (see Iliescu et al., 2020).

is greater than the US-oriented component, and when β_{US} is higher than α_{CS} the reverse is true. Individual differences in α_{CS} and β_{US} would be reflected in both CS-oriented and US-oriented responding and learning through the error-correcting learning rules. It is now time to consider how the extended associative structures depicted in **Figure 3** and generated through Equations 1–4, affect behavior.

First, we should specify how the excitatory links in the middle and right panels of **Figure 3** are integrated when either A or X is presented. When A is presented, we can assume that its associative influence (denoted $V_{CHAIN\ A\text{-}X\text{-}US}$) is the product of the numerical value of $V_{A\text{-}X}$ and $V_{COMB\ X\text{-}US}$; where $V_{COMB\ X\text{-}US}$ is calculated in the manner described in the context of combining the reciprocal associations between a CS and US. To capture the additional effect of the inhibitory link between A and the US (in the right-hand panel of **Figure 3**) the influence of $V_{COMB\ A\text{-}US}$ needs to be added. $V_{COMB\ A\text{-}US}$ has a negative value in second-order conditioning and a value of zero in sensory preconditioning (see the bracketed terms in Equations 5–7). In contrast, should X be presented, $V_{COMB\ X\text{-}US}$ would be combined with the $V_{CHAIN\ X\text{-}A\text{-}US}$.

Now, these combined values can be separated into three components that influence the links from A, X, and the US to r1-r6 in proportion to their (perceived) intensities (see Equations 5–7). Upon presentation of A at test, its intensity would be directly given (i.e., by α_A; unless one was assessing test performance during its trace; see Lin et al., 2013); while that of the (retrieved) X would be given by the absolute numerical value of $V_{A\text{-}X}$ (for sensory preconditioning), and the sum of the absolute numerical values of $V_{A\text{-}X}$ and $V_{A\text{-}US\text{-}X}$ (for second-order conditioning). This allows the perceived intensity of a retrieved stimulus to exceed its α value, in much the same way as the Rescorla-Wagner model (see Kremer, 1978). β_{US} would be given by the absolute numerical value of $V_{A\text{-}X\text{-}US}$ for sensory preconditioning, while for second-order conditioning it would be given by the absolute numerical value of the sum of $V_{A\text{-}X\text{-}US}$ + $V_{A\text{-}US}$. The fact that the link from A to the US is indirect and weak, in contrast to the direct link between X and the US, will result in a greater bias toward CS-oriented (R_A) than US-oriented (R_{US}) behaviors during A than during X (see Dwyer et al., 2012; Holland and Rescorla, 1975; Stanhope, 1992).

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

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

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

The influence of R_A, R_X, and R_{US} on the response-generating units (r1-r6 in **Figure 3**) will reflect the strengths of the unconditioned links between A, X and the US and r1-r6; for example, through multiplying R_A, R_X, and R_{US} by the weights from A, X and the US to r1-r6 (see Honey et al., 2020a). **Figure 4** presents some indicative simulations of the values of R_A, R_X, and R_{US}.

The upper panels of **Figure 4** depict simulations of sensory preconditioning, while its lower panels depict simulations of second-order conditioning. In both cases, $\alpha_A = \alpha_X = \beta_{US} = 0.80$. The left-hand panels show the values of R_A, R_X, and R_{US} for the presentation of A, which were calculated after 10 A→X trials and 2 X→US trials (sensory preconditioning) and after 10 X→US trials and 2 A→X trials (second-order conditioning). The right-hand panels show the corresponding values for the presentation of X. Values that are positive indicate the presence of higher-order conditioning. In the upper left panel, R_A and R_X output values are positive and similar, with both being higher than R_{US}. The similar output values for R_A and R_X reflect that they have the same α value and $V_{A\text{-}X}$ (the numerator in Equation 6) $\approx \alpha_X$ because it has approached asymptote over the course of 10 A→X trials. R_{US} has a lower value since the numerator in Equation 7 derives from the (absolute) numerical value of $V_{A\text{-}X} \times V_{X\text{-}US}$; which aligns to the perceived intensity of the US as retrieved by A through X. The upper right-hand panel shows the corresponding values for X[6]. R_A is lower than R_X and R_{US} because the value of $V_{X\text{-}A}$ declines over the course of X→US pairings. These simulations reveal that while R_A and R_X (aligned to CS-oriented responding) are similar whether A or X is tested, R_{US} (aligned to US-oriented responding) takes a higher value during X than A.

The lower panels of **Figure 4** show output values for simulations of second-order conditioning, generated with the same parameters as sensory preconditioning, and after the same number of trials in the first and second stages (10 X→US trials and 2 A→X[7]). Comparing first the upper and lower panels (noting their different scales), R_A and R_X output values were relatively similar during A (and X) for simulations of sensory preconditioning and second-order conditioning (see Barnet et al., 1991). However, R_{US} values were far lower for second-order conditioning than for sensory preconditioning. Indeed, if α_A and α_X are set to lower values, it results in the components of the excitatory chain becoming less effective with the consequence that there is now no second-order conditioning. In any case, the fact that R_{US} is particularly low for second-order conditioning (relative to R_A and R_X) reflects the influence of the inhibitory $V_{A\text{-}US}$ on the calculated value of β_{US}: When A is tested, the value of β_{US} = numerical values of $V_{A\text{-}US}$ (inhibitory) + $V_{A\text{-}X} \times V_{X\text{-}US}$ (excitatory); and when X is tested, β_{US} = numerical values of $V_{X\text{-}A} \times V_{A\text{-}US}$ (inhibitory) + $V_{X\text{-}US}$ (excitatory). A further difference from sensory preconditioning is that during the test with A the output value for R_A is greater than for R_X. This difference derives from the fact that in sensory preconditioning $V_{A\text{-}X}$ (the numerator in Equation 6) $\approx \alpha_X$, whereas in second-order conditioning $V_{A\text{-}X}$ does not

[6]Here, α_A = the (absolute) numerical value of $V_{X\text{-}A}$ (i.e., $1/c|V_{X\text{-}A}|$), $\alpha_X = \alpha_X$ and β_{US} = the (absolute) numerical value of $V_{X\text{-}US}$.

[7]Maintaining the number of trials of the two types (10 A→X and 2 X→US), rather than the number of trials in the two stages (10 for stage 1 and 2 for stage 2), results in extinction of the X→US association over the course of the 10 A→X trials.

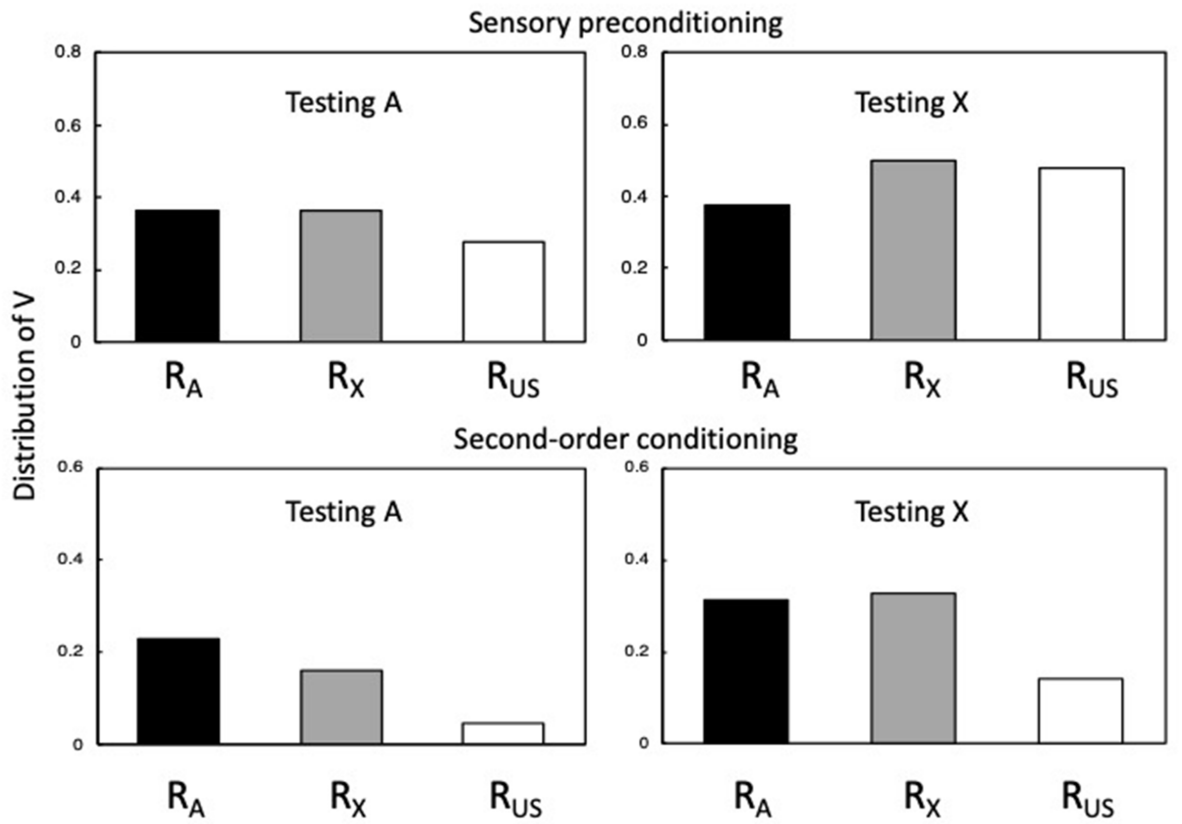

FIGURE 4 | Simulations of sensory preconditioning and second-order conditioning. The output values for R_A, R_X, and R_{US} were generated for A and X using Equations 1–7 with $\alpha_A = \alpha_X = \beta_{US} = 0.80$. There were 10 A→X trials and 2 X→US trials for the sensory preconditioning simulation, 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. Adapted from Honey and Dwyer (under review).

reach asymptote as a consequence of two A→X trials, and is further constrained by $V_{A\text{-}US\text{-}X}$ being negative. The simulations in **Figure 4** can be aligned with results reported by Stanhope (1992) using an autoshaping procedure in pigeons, and Dwyer et al. (2012) using a flavor-aversion procedure in rats: If pecking a keylight (in pigeons) and fluid consumption (in rats) is equated to CS-oriented responding (generated by R_A and R_X), and the force of pecks and lick cluster size is equated with US-oriented responding (generated by R_{US}).

Similarity Function

The central idea captured in Equations 5–7 is that the relative intensities of components of the test pattern (some present and others retrieved) determine how the associative structures depicted in **Figure 3** generate behaviors aligned to those components (A, X, and US). What they do not capture is how differences in the intensities of a given component between test and conditioning influences R_A, R_X, and R_{US}. In Equations 5–7 identity is simply assumed. There are three reasons why this needs to be addressed: First, Equations 5–7 have no (internal) mechanism for restricting conditioned behavior to stimuli that have been present on conditioning trials or to those associated with them: Associatively neutral stimuli might well influence the distribution of associative strength, but without necessarily eliciting anything other than unconditioned responses (see Pavlov, 1927, p. 44; see also, Honey et al., 2020a). Second, animals can learn discriminations in which the effective stimuli involve: (a) whether the same stimulus is presented at one intensity or a different intensity (e.g., Inman et al., 2016; for a review, see Inman and Pearce, 2018), and (b) whether the same stimulus has been presented more or less recently (e.g., Lin and Honey, 2010; see also, Pavlov, 1927; Mackintosh, 1974, p. 104; Staddon and Higa, 1999; Staddon, 2005). The latter observation reducing to the former once different components of a decaying trace are equated with different stimulus intensities; both observations suggest that different intensities of a given stimulus can enter into different associations, but also that there is generalization between those intensities. Third, the idea that the representation of the CS includes the intensity at which it is presented affords an account for when higher-order conditioning is observed: As already noted, trace conditioning might enhance higher-order conditioning because when A retrieves X at test (i.e., X*) it is more similar in perceived intensity to the stimulus that became linked to the US during trace conditioning (X*) than

standard conditioning (X; Ward-Robinson and Hall, 1998; Lin and Honey, 2011; see also, Kamil, 1969; Cole et al., 1995; Barnet and Miller, 1996). It would also help to explain the fact that higher-order conditioning to A can be left unaffected by the extinction of responding to X (e.g., Rizley and Rescorla, 1972; Cheatle and Rudy, 1978; Amiro and Bitterman, 1980; Nairne and Rescorla, 1981; Archer and Sjödén, 1982; Ward-Robinson and Hall, 1996; but see, Rescorla, 1982): Because X (rather than the trace, X*) would undergo extinction when X is presented (see Kamin, 1969; Mackintosh, 1976). Finally, when A is presented with X at test, A will retrieve X*, which has strength independently of X itself (e.g., Ward-Robinson et al., 2001; Lin et al., 2013). This analysis is plausible, but without a function that specifies the similarity between the perceived intensities of stimuli, their traces, and retrieved representations it remains tendentious (see Lin and Honey, 2011, 2016; see also, Lin et al., 2013). However, one such function is presented below in the context of how the retrieved memory of X affects performance during the presentation of A (i.e., in a modification of Equation 6).

$$R_X = \frac{\alpha_X}{\alpha_A + \alpha_X + \beta_{US}}\left(\left(\alpha_{X\text{-}R}S\alpha_{X\text{-}C} \times V_{CHAIN\ A\text{-}X\text{-}US}\right) + V_{COMB\ A\text{-}US}\right)$$

Where: (8)

$$\alpha_X = \alpha_{X\text{-}R} = \left|\tfrac{1}{C}V_{A\text{-}X}\right| \text{ and } \alpha_{X\text{-}C} = \alpha \text{ of X upon delivery of the US}$$

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

The function ($\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$) introduced in Equation 8 (in the gray boxes) determines the similarity (S) of two values: The numerical value of $V_{A\text{-}X}$ (denoted $\alpha_{X\text{-}R}$) and its conditioned counterpart or trace (denoted $\alpha_{X\text{-}C}$). It is worth remembering that when $V_{A\text{-}X}$ reaches asymptote, its numerical value $\approx \alpha_X$, which means that $\alpha_{X\text{-}R} \approx \alpha_{X\text{-}C}$. This function is also applied to modify the bracketed term in Equations 5 and 7 when A is presented. Its basic properties are simple: When the values of $\alpha_{X\text{-}R}$ and $\alpha_{X\text{-}C}$ are close together then $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ approaches 1, but as they diverge then $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ approaches 0. Applying these ideas to how $\alpha_{X\text{-}R}$ affects performance is also simple. Because the asymptote for $V_{A\text{-}X}$ during A→X training is α_X, when A is presented at test $\alpha_{X\text{-}R}$ will have approached α_X over the A→X trials. If A→X training had proceeded until $V_{A\text{-}X}$ reached asymptote then $\alpha_{X\text{-}R}$ and $\alpha_{X\text{-}C}$ would be maximally similar, provided α_X during X→US conditioning trials was the same as during A→X trials (as it usually is). Now, we can appreciate how $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ varies when α_X has one value for A→X trials (e.g., 0.50) and is then reduced for X→US trials (e.g., 0.45); this reduction in $\alpha_{X\text{-}C}$ is intended to mimic the effect of introducing a trace interval between X and the US (see Lin and Honey, 2011, 2016; Lin et al., 2013). It should be clear that before $V_{A\text{-}X}$ has reached asymptote during A→X trials, its numerical value can match more closely 0.45 than 0.50; and that as $V_{A\text{-}X}$ tends to 0.50 for A→X trials the numerical value of $V_{A\text{-}X}$ will become closer to 0.50 than 0.45. The accuracy of this analysis was confirmed by simulations.

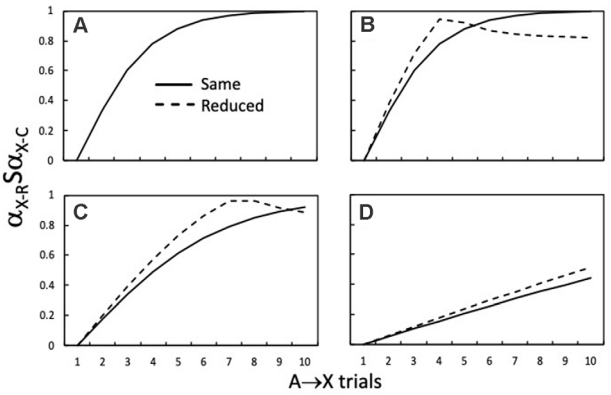

FIGURE 5 | How the similarity ($\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$) of the retrieved X ($\alpha_{X\text{-}R}$) to the conditioned X ($\alpha_{X\text{-}C}$) during a test with A varies with the number of initial A→X trials. The continuous lines denote $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ output values when the α_X value (0.50) used to compute changes in $V_{A\text{-}X}$ (i.e., $\alpha_{X\text{-}R}$) was the *same* as that for $\alpha_{X\text{-}C}$ on X→US conditioning trials. α_A was 0.50 in panels **(A,B)**, 0.30 in panel **(C)**, and 0.10 in panel **(D)**. The dashed lines denote $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ output values when the α_X value used to compute changes in $V_{A\text{-}X}$ (0.50; i.e., $\alpha_{X\text{-}R}$) was *reduced* to 0.45 for $\alpha_{X\text{-}C}$ to calculate $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$. This manipulation is akin to using trace conditioning for X→US trials. Adapted from Honey and Dwyer (under review).

Figure 5 shows how $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ varies as a function of the number of A→X training trials during the first stage of training. The continuous lines show $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ when $\alpha_{X\text{-}R}$ and $\alpha_{X\text{-}C}$ are generated by the same α_X value (e.g., 0.50), as in standard higher-order conditioning procedures. Comparison of the continuous lines across **Figures 5A–D** shows that the rate at which maximum similarity is approached, across a series of A→X trials, decreases as α_A is reduced from 0.50 (**Figures 5A,B**), to 0.30 (**Figure 5C**), and then 0.10 (**Figure 5D**). Turning now to the dashed lines in **Figures 5B–D**, it is clear that there is a period of initial A→X training when reductions in $\alpha_{X\text{-}C}$ increase $\alpha_{X\text{-}R}S\alpha_{X\text{-}C}$ compared to when $\alpha_{X\text{-}C}$ is the same (i.e., 0.50 for the continuous lines). With more extended A→X training this pattern reverses as $\alpha_{X\text{-}R}$ (i.e., $V_{A\text{-}X}$) approaches 0.50 and consequently deviates from the reduced value of $\alpha_{X\text{-}C}$ (i.e., 0.45). This reversal is apparent in **Figures 5B,C**, but not within 10 trials in **Figure 5D**.

The take-home message from these simulations is that trace conditioning will have the potential to enhance higher-order conditioning if A is tested when A→X training has left $V_{A\text{-}X}$ within the range where the dashed line has higher values than the continuous line. The influence of such increases in similarity on higher-order conditioning will be contingent on them more than counteracting any direct effect of reducing $\alpha_{X\text{-}C}$ on the efficacy of the X-US component of the chain (i.e., $V_{CHAIN\ A\text{-}X\text{-}US}$). In fact, simulations reveal that increases in R_A, R_X, and R_{US} of between 10% to 20% are produced by reducing $\alpha_{X\text{-}C}$ by 10%, which is in the range where reducing $\alpha_{X\text{-}C}$ has little effect on the rate at which $V_{X\text{-}US}$ approaches the asymptote determined by β_{US}. These effects of similarity are more marked for sensory preconditioning

than second-order conditioning (see Honey and Dwyer, under review).

Our formal analysis assumes that the value of α on a conditioning trial is encoded and is one basis for generalization between a CS presented at one intensity and the same CS but delivered at a different intensity. It also assumes that there is a (computational) equivalence between different α (and β) values generated by changing a stimulus physically (e.g., Inman et al., 2016) and the values generated through the central processes of decay and retrieval (e.g., Lin and Honey, 2010; see also Iliescu et al., 2020). In addition to providing an analysis for how trace conditioning can enhance higher-order conditioning, it can also explain related observations: the facts that extinction of X is not always reflected in responding to A, and the compound AX generates more responding than X in sensory preconditioning procedures. As already noted, the effects of extinction treatments involving the presentation of X will be more likely to impact its directly activated α value as opposed to its decaying value through a process of overshadowing (Mackintosh, 1976); and whether this affects higher-order conditioning will depend on whether the representation of X that supports responding to A (which is determined by the strength of the A→X association; see Rescorla, 1982) is similar to its directly activated or decaying forms. Equation 8 provides a formal example of how test performance is affected by the similarity between the value of X retrieved by A as a consequence of A→X trials and its encoded value during conditioning trials. According to our analysis, AX will generate more responding than X because the associative chain can exert an independent influence on the US representation (for further details, see Honey and Dwyer, under review).

To close the theoretical loop, the learning rules (e.g., $\Delta V_{1\text{-}2} = \alpha_1(c.\alpha_2 - \Sigma V_{S\text{-}TOTAL\text{-}2})$) can be modified to reflect the fact that the associative strengths of stimuli contributing to $\Sigma V_{S\text{-}TOTAL\text{-}2}$ (including $V_{1\text{-}2}$) need to be scaled by their similarity (subscript s) to their intensities when conditioned (see Pearce, 1994). For instance, Equation 3 can be re-cast as Equation 9, where the subscript s denotes this scaling process. The similarity function is as before, but $\alpha_{X\text{-}R}$ is the perceived intensity of the CS on previous trials, while $\alpha_{X\text{-}C} = \alpha_X$ of the same CS on the current trial. In this way, the perceived intensity of a CS is encoded as one component of what is learned on a conditioning trial (if α_X changes from one trial to the next then new learning occurs), which reflects the generalization of associative strength between a stimulus conditioned at one intensity and later presented at another intensity (i.e., $\Sigma V_{S\text{-}TOTAL\text{-}US}$ is reduced because $\alpha_{X\text{-}R}S\alpha_{X\text{-}C} < 1$). It should be recognized, however, that increases and reductions in intensity have different effects on behavior through the proportion terms in the equations that determine the distribution of associative strength (e.g., in Equation 8). Finally, it is worth noting that the effect of changing α_X from one trial to the next on the US→X association will be that $V_{US\text{-}X}$ homes in on the new α_X (see Equation 4), which parallels the fact that changes in US intensity across trials affects the asymptote of the X-US association.

$$\Delta V_{X\text{-}US} = \alpha_X(c.\beta_{US} - \Sigma V_{S\text{-}TOTAL\text{-}US}) \quad (9)$$

DISCUSSION: SOME CONCLUDING CONSIDERATIONS

Understanding higher-order conditioning has theoretical and translational value, but traditional (informal) accounts of this phenomenon are poorly equipped to address two fundamental issues: What is learned and how it is expressed. The analysis described here and developed in Honey and Dwyer (under review) borrows from HeiDI, which is a model of Pavlovian learning and performance (Honey et al., 2020a). The learning and performance rules are derived from HeiDI, but their influence is modulated by a similarity function. This function specifies the similarity between the same nominal stimulus, which can take different perceived intensities as a result of manipulating the intensity at which it is delivered and through processes of retrieval or trace decay. The resulting analysis has clear implications for behavioral neuroscience, where group-level differences in higher-order conditioning should be interpreted with caution: Changes in a given behavioral measure of higher-order conditioning consequent on a manipulation might have a variety of origins. For example, differences in learning or performance might not reflect differences in the underlying learning mechanisms but rather changes to: α (for A and X), β (for the US), or their associated decay functions (see Honey and Good, 2000); or indeed the requisite (neural) computations involving the processes represented by these parameters.

In developing this more formal analysis of higher-order conditioning, no appeal has been made to any process of retrieval mediated learning or stimulus-response learning. This is not intended to suggest that such forms of learning are without consequence, but simply that they are not required by the available evidence. For example, the model presented here could accommodate retrieval mediated learning between A and the US in a sensory preconditioning procedure by substituting the numerical value of $\Sigma V_{TOTAL\text{-}A}$ for α_A: $\Delta V_{A\text{-}US} = 1/c.\Sigma V_{TOTAL\text{-}A}(c.\beta_{US} - \Sigma V_{S\text{-}TOTAL\text{-}US})$; recall that multiplying $\Sigma V_{TOTAL\text{-}A}$ by 1/c transforms it into a dimensionless scalar like α_A. In this way, a retrieved stimulus, or stimulus trace, might acquire associative strength while limiting that acquired by other stimuli present on a conditioning trial. As we have noted, retrieved stimuli will also affect performance through the proportion terms in Equations 5–8 (see Holland, 1983). This analysis joins others that have attempted to provide a more specific account of the process of retrieval mediated learning, albeit that they do not apply as readily to higher-order conditioning as they do to other phenomena (e.g., Van Hamme and Wasserman, 1994; Dickinson and Burke, 1996; see also, Dwyer et al., 1998).

We should briefly comment on the complexity of the model. While the model has three components (relating to learning, performance, and similarity) it only has two free parameters: α (for A and X) and β (for the US); and their associated decay functions. It can also be summarized in two simple statements: 1. The perceived intensities of stimuli present during a test affect how learning represented within an extended associative structure affects performance; and 2. The similarity of the

perceived intensities of the tested stimuli to conditioned stimuli within that structure modulates the translation of learning into performance.

Our use of the term *perceived intensity* clearly affords a potential analysis of individual differences in both Pavlovian conditioning and higher-order conditioning at the level of learning and performance (see Honey et al., 2020a,b,c), but also now in terms of the similarity between directly activated representations, their decaying traces, and retrieved forms. Pavlov (1927; p. 105) noted that there were marked individual differences in the strength of second-order reflexes: "*Among the experimental dogs one finds special types of nervous systems; in particular there are dogs with weak nervous systems in which this phenomenon is clearly expressed.*" The fact that there are significant individual differences in how learning is evident in behavior has been neglected by general-process models of learning. The model upon which our analysis is based, HeiDI, represents a prosaic approach to accommodating both quantitative and qualitative individual differences in conditioned behavior.

AUTHOR CONTRIBUTIONS

RH and DD contributed to the ideas presented in this article and to its preparation for publication. Both authors contributed to the article and approved the submitted version.

FUNDING

The development of the ideas, together with our research that informed them, was supported by a BBSRC grant (UK; BB/T004339/1; PI: RH).

ACKNOWLEDGMENTS

We thank A. F. Iliescu for comments on a draft of this article.

REFERENCES

Allman, M. J., and Honey, R. C. (2006). Transfer of configural learning between the components of a preexposed stimulus compound: implications for elemental and configural models of learning. *J. Exp. Psychol. Anim. Behav. Process.* 32, 307–313. doi: 10.1037/0097-7403.32.3.307

Amiro, T. W., and Bitterman, M. E. (1980). Second-order appetitive conditioning in goldfish. *J. Exp. Psychol. Anim. Behav. Process.* 6, 41–48. doi: 10.1037/0097-7403.6.1.41

Arcediano, F., Escobar, M., and Miller, R. R. (2005). Bidirectional associations in humans and rats. *J. Exp. Psychol. Anim. Behav. Process.* 31, 301–318. doi: 10.1037/0097-7403.31.3.301

Archer, T., and Sjödén, P. (1982). Higher-order conditioning and sensory preconditioning of a taste aversion with an exteroreceptive CS1. *Q. J. Exp. Psychol. B* 34, 1–17. doi: 10.1080/14640748208400886

Asch, S., and Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proc. Am. Philos. Soc.* 106, 135–163.

Asratian, E. A. (1965). *Compensatory Adaptations, Reflex Activity, and the Brain.* Oxford: Pergamon Press.

Barnet, R. C., Grahame, N. J., and Miller, R. R. (1991). Comparing the magnitudes of second-order conditioning and sensory preconditioning effects. *Bull. Psychon. Soc.* 29, 133–135. doi: 10.3758/bf03335215

Barnet, R. C., and Miller, R. R. (1996). Second-order excitation mediated by a backward conditioned inhibitor. *J. Exp. Psychol. Anim. Behav. Process.* 22, 279–296. doi: 10.1037//0097-7403.22.3.279

Boakes, R. A. (1977). "Performance on learning to associate a stimulus with positive reinforcement," in *Operant-Pavlovian Interactions*, eds H. Davis and H. M. B. Hurwitz (Hillsdale, NJ: Lawrence Erlbaum Associates), 67–101.

Brogden, W. J. (1939). Sensory pre-conditioning. *J. Exp. Psychol.* 25, 323–332. doi: 10.1037/h0058944

Cheatle, M. D., and Rudy, J. W. (1978). Analysis of second-order odor-aversion conditioning in neonatal rats: implications for Kamin's blocking effect. *J. Exp. Psychol. Anim. Behav. Process.* 4, 237–249. doi: 10.1037//0097-7403.4.3.237

Cole, R. P., Barnet, R. C., and Miller, R. R. (1995). Temporal encoding in trace conditioning. *Anim. Learn. Behav.* 23, 144–153. doi: 10.3758/bf03199929

Cole, R. P., and Miller, R. R. (1999). Conditioned excitation and conditioned inhibition acquired through backward conditioning. *Learn. Motiv.* 30, 129–156. doi: 10.1006/lmot.1998.1027

Crawford, L. L., and Domjan, M. (1995). Second-order sexual conditioning in male Japanese quail (*Coturnix japonica*). *Anim. Learn. Behav.* 23, 327–334. doi: 10.3758/bf03198929

Davey, G. C., and Cleland, G. G. (1982). Topography of signal-centred behavior in the rat: effects of deprivation state and reinforcer type. *J. Exp. Anal Behav.* 38, 291–304. doi: 10.1901/jeab.1982.38-291

Davey, G. C. L., and Arulampalan, T. (1982). Second-order "fear" conditioning in humans. Persistence of CR2 following extinction of CR1. *Behav. Res. Ther.* 20, 391–396. doi: 10.1016/0005-7967(82)90099-7

Davey, G. C. L., and McKenna, I. (1983). The effects of post-conditioning revaluation of CS1 and UCS following Pavlovian second-order electrodermal conditioning in humans. *Q. J. Exp. Psychol. B* 35, 125–133. doi: 10.1080/14640748308400899

Dickinson, A., and Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Q. J. Exp. Psychol. B* 49, 60–80. doi: 10.1080/713932614

Dwyer, D. M. (2012). Licking and liking: the assessment of hedonic responses in rodents. *Q. J. Exp. Psychol.* 65, 371–394. doi: 10.1080/17470218.2011.652969

Dwyer, D. M., Burgess, K. V., and Honey, R. C. (2012). Avoidance but not aversion following sensory-preconditioning with flavors: a challenge to stimulus substitution. *J. Exp. Psychol. Anim. Behav. Process.* 38, 359–368. doi: 10.1037/a0029784

Dwyer, D. M., Mackintosh, N. J., and Boakes, R. A. (1998). Simultaneous activation of the representations of absent cues results in the formation of an excitatory association between them. *J. Exp. Psychol. Anim. Behav. Process.* 24, 163–171. doi: 10.1037/0097-7403.24.2.163

Field, A. P. (2006). Is conditioning a useful framework for understanding the development and treatment of phobias? *Clin. Psychol. Rev.* 26, 857–875. doi: 10.1016/j.cpr.2005.05.010

Flagel, S. B., Akil, H., and Robinson, T. E. (2009). Individual differences in the attribution of incentive salience to reward-related cues: implications for addiction. *Neuropharmacology* 56, 139–148. doi: 10.1016/j.neuropharm.2008.06.027

Gerolin, M., and Matute, H. (1999). Bidirectional associations. *Anim. Learn. Behav.* 27, 42–49. doi: 10.3758/BF03199430

Gewirtz, J. C., and Davis, M. (2000). Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. *Learn. Mem.* 7, 257–266. doi: 10.1101/lm.35200

Gilboa, A., Sekeres, M., Moskovitch, M., and Winocur, G. (2014). Higher-order conditioning is impaired by hippocampal lesions. *Curr. Biol.* 24, 2202–2207. doi: 10.1016/j.cub.2014.07.078

Hall, G. (1996). Learning about associatively activated stimulus representations: implications for acquired equivalence and perceptual learning. *Anim. Learn. Behav.* 24, 233–255. doi: 10.3758/bf03198973

Haselgrove, M., and Hogarth, L. (2011). *Clinical Applications of Learning Theory*. UK: Psychology Press.

Hearst, E., and Jenkins, H. (1974). *Sign-Tracking: The Stimulus-Reinforcer Relation and Directed Action.* Austin, TX: Monograph of the Psychonomic Society.

Hebb, D. O. (1949). *The Organization of Behavior*. New York, NY: Wiley and Sons.

Heth, C. D. (1976). Simultaneous and backward fear conditioning as a function of number of CS-UCS pairings. *J. Exp. Psychol. Anim. Behav. Process.* 2, 117–129. doi: 10.1037//0097-7403.2.2.117

Holland, P. C. (1980). Second-order conditioning with and without the US. *J. Exp. Psychol. Anim. Behav. Process.* 6, 238–250. doi: 10.1037//0097-7403.6.3.238

Holland, P. C. (1981). Acquisition of a representation-mediated conditioned food aversion. *Learn. Motiv.* 12, 1–18. doi: 10.1016/0023-9690(81)90022-9

Holland, P. C. (1983). Representation-mediated overshadowing and potentiation of conditioned aversions. *J. Exp. Psychol. Anim. Behav. Process.* 9, 1–13. doi: 10.1037/0097-7403.9.1.1

Holland, P. C. (1984). Origins of behavior in Pavlovian conditioning. *Psychol. Learn. Motiv.* 18, 129–174. doi: 10.1016/s0079-7421(08)60361-8

Holland, P. C. (1977). Conditioned stimulus as a determinant of the form of the Pavlovian conditioned response. *J. Exp. Psychol. Anim. Behav. Process.* 3, 77–104. doi: 10.1037//0097-7403.3.1.77

Holland, P. C. (2016). Enhancing second-order conditioning with lesions of the basolateral amygdala. *Behav. Neurosci.* 130, 176–181. doi: 10.1037/bne0000129

Holland, P. C., and Rescorla, R. A. (1975). Second-order conditioning with food unconditioned stimulus. *J. Comp. Physiol. Psychol.* 88, 459–467. doi: 10.1037/h0076219

Honey, R. C., Dwyer, D. M., and Iliescu, A. F. (2020a). HeiDI: a model for Pavlovian learning and performance with reciprocal associations. *Psychol. Rev.* 127, 829–852. doi: 10.1037/rev0000196

Honey, R. C., Dwyer, D. M., and Iliescu, A. F. (2020b). Elaboration of a model of Pavlovian learning and performance: HeiDI.. *J. Exp. Psychol. Anim. Learn. Cogn.* 46, 170–184. doi: 10.1037/xan0000239

Honey, R. C., Dwyer, D. M., and Iliescu, A. F. (2020c). Individual variation in the vigor and form of Pavlovian conditioned responses: analysis of a model system. *Learn. Motiv.* 72:101658. doi: 10.1016/j.lmot.2020.101658

Honey, R. C., and Good, M. (2000). Associative components of recognition memory. *Curr. Opin. Neurobiol.* 10, 200–204. doi: 10.1016/s0959-4388(00)00069-6

Honey, R. C., Good, M., and Manser, K. L. (1998a). Negative priming in associative learning: evidence from a serial-habituation procedure. *J. Exp. Psychol. Anim. Behav. Process.* 24, 229–237. doi: 10.1037//0097-7403.19.1.90

Honey, R. C., Watt, A., and Good, M. (1998b). Hippocampal lesions disrupt an associative mismatch process. *J. Neurosci.* 18, 2226–2230. doi: 10.1523/JNEUROSCI.18-06-02226.1998

Hull, C. L. (1949). Stimulus intensity dynamism (V) and stimulus generalization. *Psychol. Rev.* 56, 67–76. doi: 10.1037/h0058051

Iliescu, A. F., Dwyer, D. M., and Honey, R. C. (2020). Individual differences in the nature of conditioned behavior across a conditioned stimulus: adaptation and application of a model. *J. Exp. Psychol. Anim. Learn. Cogn.* 46, 460–469. doi: 10.1037/xan0000270

Iliescu, A. F., Hall, J., Wilkinson, L., Dwyer, D. M., and Honey, R. C. (2018). The nature of phenotypic variation in Pavlovian conditioning. *J. Exp. Psychol. Anim. Learn. Cogn.* 44, 358–369. doi: 10.1037/xan0000177

Inman, R. A., Honey, R. C., and Pearce, J. M. (2016). "Asymmetry in the discrimination of auditory intensity: implications for theories of stimulus generalisation," in *Associative Learning and Cognition. Homage to Professor N.J. Mackintosh*, eds J. B. Trobalon and V. D. Chamizo (Barcelona: Edicions de la Universitat de Barcelona), 197–222.

Inman, R. A., and Pearce, J. M. (2018). The discrimination of magnitude: a review and theoretical analysis. *Neurobiol. Learn. Mem.* 153, 118–130. doi: 10.1016/j.nlm.2018.03.020

Iordanova, M. D., Good, M., and Honey, R. C. (2011). Retrieval-mediated learning involving episodes requires synaptic plasticity in the hippocampus. *J. Neurosci.* 31, 7156–7162. doi: 10.1523/JNEUROSCI.0295-11.2011

Jenkins, H., and Moore, B. R. (1973). The form of the autoshaped response with food or water reinforcer. *J. Exp. Anal. Behav.* 20, 163–181. doi: 10.1901/jeab.1973.20-163

Kamil, A. C. (1969). Some parameters of the second-order conditioning of fear in rats. *J. Comp. Physiol. Psychol.* 67, 364–369. doi: 10.1037/h0026782

Kamin, L. (1969). "Selective association and conditioning," in *Fundamental Issues in Associative Learning*, eds N. J. Mackintosh and W. K. Honig (Halifax: Dalhousie University Press), 42–89.

Konorski, J. (1948). *Conditioned Reflexes and Neuron Organization.* Cambridge, MA: Cambridge University Press.

Kremer, E. F. (1978). The Rescorla-Wagner model: losses in associative strength in compound conditioned stimuli. *J. Exp. Psychol. Anim. Behav. Process.* 4, 22–36. doi: 10.1037//0097-7403.4.1.22

Lay, B. P. P., Westbrook, R. F., Glanzman, D. L., and Holmes, N. N. (2018). Commonalities and differences in the substrates underlying consolidation of first- and second-order conditioned fear. *J. Neurosci.* 38, 1926–1941. doi: 10.1523/JNEUROSCI.2966-17.2018

Lin, T.-E., Dumigan, N. M., Dwyer, D. M., Good, M. A., and Honey, R. C. (2013). Assessing the encoding specificity of associations with sensory preconditioning procedures. *J. Exp. Psychol. Anim. Behav. Process.* 39, 67–75. doi: 10.1037/a0030662

Lin, T.-E., Dumigan, N. M., Good, M. A., and Honey, R. C. (2016). Novel sensory preconditioning procedures identify a specific role for the hippocampus in pattern completion. *Neurobiol. Learn. Mem.* 130, 142–148. doi: 10.1016/j.nlm.2016.02.006

Lin, T.-E., and Honey, R. C. (2010). Analysis of the content of configural representations: the role of associatively evoked and trace memories. *J. Exp. Psychol. Anim. Behav. Process.* 36, 501–505. doi: 10.1037/a0018348

Lin, T. E., and Honey, R. C. (2011). Encoding specific associative memory: evidence from behavioral and neural manipulations. *J. Exp. Psychol. Anim. Behav. Process.* 37, 317–329. doi: 10.1037/a0022497

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

Mackintosh, N. J. (1974). *The Psychology of Animal Learning.* London: Academic Press.

Mackintosh, N. J. (1975). A theory of attention: variations in the associability of stimuli with reinforcement. *Psychol. Rev.* 82, 276–298. doi: 10.1037/h0076778

Mackintosh, N. J. (1976). Overshadowing and stimulus intensity. *Anim. Learn. Behav.* 4, 186–192. doi: 10.3758/bf03214033

Mackintosh, N. J. (1983). *Conditioning and Associative Learning.* Oxford: Oxford University Press.

Maes, E. J. P., Sharpe, M. J., Usypchuk, A., Lozzi, M., Gardner, M. P. H., Chang, C. Y., et al. (2020). Causal evidence supporting the proposal that dopamine transients function as temporal difference prediction errors. *Nat. Neurosci.* 23, 176–178. doi: 10.1038/s41593-019-0574-1

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

Miller, R. R., and Barnet, R. C. (1993). The role of time in elementary associations. *Curr. Direct. Psychol. Sci.* 2, 106–111. doi: 10.1111/1467-8721.ep10772577

Mollick, J. A., Hazy, T. E., Krueger, K. A., Nair, A., Mackie, P., Herd, S. E., et al. (2020). A systems neuroscience model of phasic dopamine. *Psychol. Rev.* 6, 972–1021. doi: 10.1037/rev0000199

Nairne, J. S., and Rescorla, R. A. (1981). Second-order conditioning with diffuse auditory reinforcers in the pigeon. *Learn. Motiv.* 12, 65–91doi: 10.1016/0023-9690(81)90025-4

Patitucci, E., Nelson, A. J. D., Dwyer, D. M., and Honey, R. C. (2016). The origins of individual differences in how learning is expressed in rats: a general-process perspective. *J. Exp. Psychol. Anim. Learn. Cogn.* 42, 313–324. doi: 10.1037/xan0000116

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

Pearce, J. M. (1994). Similarity and discrimination: a selective review and a connectionist model. *Psychol. Rev.* 101, 587–607. doi: 10.1037/0033-295x.101.4.587

Pearce, J. M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not unconditioned stimuli. *Psychol. Rev.* 87, 532–552. doi: 10.1037/0033-295X.87.6.532

Rashotte, M., Griffin, R. W., and Sisk, C. L. (1977). Second-order conditioning of the pigeon's keypeck. *Anim. Learn. Behav.* 5, 25–38. doi: 10.3758/bf03209127

Rescorla, R. A. (1982). Simultaneous second-order conditioning produces S-S learning in conditioned suppression. *J. Exp. Psychol. Anim. Behav. Process.* 8, 23–32. doi: 10.1037/0097-7403.8.1.23

Rescorla, R. A., and Cunningham, C. L. (1978). Within-compound flavor associations. *J. Exp. Psychol. Anim. Behav. Process.* 4, 267–275. doi: 10.1037//0097-7403.4.3.267

Rescorla, R. A., and Wagner, A. R. (1972). "A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. H. Black and W. F. Prokasy (New York, NY: Appleton-Century-Crofts), 64–99.

Rizley, R. C., and Rescorla, R. A. (1972). Associations in second-order conditioning and sensory preconditioning. *J. Comp. Physiol. Psychol.* 81, 1–11. doi: 10.1037/h0033333

Silva, A. I., Haddon, J. E., Trent, S., Syed, Y., Lin, T.-C. E., Patel, Y., et al. (2019). Cyfip1 haploinsufficiency is associated with white matter changes, myelin thinning, reduction of mature oligodendrocytes and behavioural inflexibility. *Nat. Commun.* 10:3455. doi: 10.1038/s41467-019-11119-7

Staddon, J. E. (2005). Interval timing: memory, not a clock. *Trends Cogn. Sci.* 9, 312–314. doi: 10.1016/j.tics.2005.05.013

Staddon, J. E. R., and Higa, J. J. (1999). Time and memory: towards a pacemaker-free theory of interval timing. *J. Exp. Anal. Behav.* 71, 215–251. doi: 10.1901/jeab.1999.71-215

Stanhope, K. J. (1992). The representation of the reinforcer and the force of the pigeon's keypeck in first- and second-order conditioning. *Q. J. Exp. Psychol. B* 44, 137–158. doi: 10.1080/02724999208250607

Tait, R. W., and Saladin, M. E. (1986). Concurrent development of excitatory and inhibitory associations during backward conditioning. *Anim. Learn. Behav.* 14, 133–137. doi: 10.3758/bf03200047

Timberlake, W., and Grant, D. L. (1975). Auto-Shaping in rats to the presentation of another rat predicting food. *Science* 190, 690–692. doi: 10.1126/science.190.4215.690

Van Hamme, L. J., and Wasserman, E. A. (1994). Cue competition in causality judgments: the role of nonpresentation of compound stimulus elements. *Learn. Motiv.* 25, 127–151. doi: 10.1006/lmot.1994.1008

Wagner, A. R. (1981). "SOP: A model of automatic memory processing in animal behavior," in *Information Processing in Animals: Memory Mechanisms*, eds N. E. Spear and R. R. Miller (Hillsdale, NJ: Erlbaum), 5–48.

Ward-Robinson, J., Coutureau, E., Good, M., Honey, R. C., Killcross, A. S., and Oswald, C. J. P. (2001). Excitotoxic lesions of the hippocampus leaves sensory preconditioning intact: implications for models of hippocampal function. *Behav. Neurosci.* 115, 1357–1362. doi: 10.1037//0735-7044.115.6.1357

Ward-Robinson, J., and Hall, G. (1996). Backward sensory preconditioning. *J. Exp. Psychol. Anim. Behav. Process.* 22, 395–404. doi: 10.1037/0097-7403.22.4.395

Ward-Robinson, J., and Hall, G. (1998). Backward sensory preconditioning when reinforcement is delayed. *Q. J. Exp. Psychol. B* 51, 349–362. doi: 10.1080/713932687

Wessa, M., and Flor, H. (2007). Failure of extinction of fear responses in post-traumatic stress disorder: evidence from second-order conditioning. *Am. J. Psychiatry* 164, 1684–1692. doi: 10.1176/appi.ajp.2007.07030525

Zentall, T. R., Sherburne, L. M., and Steirn, J. N. (1992). Development of excitatory backward associations during the establishment of forward associations in a delayed conditional discrimination by pigeons. *Anim. Learn. Behav.* 20, 199–206. doi: 10.3758/bf03213373

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.