Thinking of Learning Phenomena as Instances of Relational Behavior

Jan De Houwer, Martin Finn, Matthias Raemaekers, Jamie Cummins, and Yannick Boddez

Ghent University

In press. Learning & Behavior.

Authors Note

JDH, MF, MR, JC, YB, Department of Experimental Clinical and Health Psychology, Ghent University. The preparation of this paper was supported by Grant BOF16/MET_V/002 of Ghent University to JDH. Several of the ideas put forward in this paper were shaped during conversations with Dermot Barnes-Holmes and Sean Hughes. Correspondence concerning this paper can be sent to jan.dehouwer@ugent.be .

Abstract

We explore the idea that some learning phenomena can be thought of as instances of relational behavior, more specifically arbitrarily applicable relational responding (AARR). After explaining the nature of AARR, we discuss what it means to say that learning phenomena such as evaluative and fear conditioning are instances of AARR. We then list several implications of this perspective for empirical and theoretical research on learning, as well as for how learning phenomena relate to other psychological phenomena in human and non-human animals.

Keywords: learning, conditioning, behavior, conceptual analysis

Thinking of Learning Phenomena as Instances of Relational Behavior

For more than 100 years, psychologists have examined a host of learning phenomena such as classical and operant conditioning (see Bouton, 2016; Catania, 2013; De Houwer & Hughes, 2020, for reviews). Typically, the focus is on how events that occur during a lab-based experimental procedure change the responses of an organism. For instance, studies on evaluative conditioning in humans might involve multiple trials in which a neutral brand name (conditional stimulus; CS) is presented together with a picture of smiling people (unconditional stimulus; US). Researchers such as the first author of this paper have spent many years examining whether those CS-US pairings change evaluative responses to the CS (i.e., the evaluative conditioning effect), the moderators of this effect (e.g., the number of CS-US pairings), and the mental mechanisms via which CS-US pairings influence responses to the CS (e.g., the formation of associations in memory; see Moran et al., in press, for a review). A learning phenomenon like evaluative conditioning is thus typically considered to be a *functional* process (i.e., evaluative responses are a function of CS-US pairings) that is mediated by a *mental* process (e.g., CS-US pairings are assumed to influence evaluative responses via the formation of associations in memory; De Houwer, 2007; De Houwer et al., 2013).

In this paper, we explore a radically different perspective on learning phenomena, namely the idea that they can be thought of as part of a *behavioral* process, that is, the unfolding of a behavior that has been learned in the past and is performed in the current situation. More specifically, learning phenomena can be conceived of as instances of a type of behavior known as arbitrarily applicable relational responding (AARR). In essence, this is symbolic behavior that involves acting as if events are related in a certain way, irrespective of their physical properties. For instance, people can act as if the word GLASS is in some respects equivalent to the object glass (even though physically, there is no resemblance

between the word and the object) or that a dime is more valuable than a nickel (even though the dime is "less than" the nickel in terms of physical size). Proponents of Relational Frame Theory (RFT; e.g., Hayes et al., 2001; see Hughes & Barnes-Holmes, 2016, and Barnes-Holmes & Harte, 2022, for reviews) have argued that most cognitive and language abilities in human adults can be conceived of in terms of AARR. In line with RFT, we highlight the possibility that in studies on learning in human adults, certain events may function as cues for acting as if stimuli are related, which would result in learning effects. For instance, in evaluative conditioning studies, the fact that a neutral brand name (CS) is paired with a positive picture (US) might function as a cue for responding as if both stimuli are equivalent, which includes responding to the neutral brand name as positive (De Houwer & Hughes, 2016; Hughes et al., 2016a). From this perspective, evaluative conditioning (i.e., the change in evaluative responses to the CS that results from the CS-US pairings) occurs because people act as if the CS and US are equivalent in certain ways, a behavior that is prompted by the fact that CS and US occur together in space and time.

In the remainder of the paper, we first explain in more detail what AARR entails (see also De Houwer & Hughes, 2020, Section 3.2.5.5; Stewart & McElwee, 2009) and what it means to say that learning phenomena can be instances of AARR (see also, De Houwer & Hughes, 2017, 2020, Section 4.2). Afterwards, we discuss a number of unique implications of this idea for future empirical and theoretical research on learning (i.e., novel empirical predictions, challenges for computational models, and links with propositional theories of learning) and for conceptualizing the relation between learning phenomena in humans and other psychological phenomena (i.e., in terms of the nature of the events that function as cues for AARR or the apparent absence of fully-fledged AARR in non-human animals). These implications are derived primarily from the fact that our perspective (a) strongly emphasizes the role of events that occur before the start of a lab-based learning procedure (i.e., events that

allow for the acquisition of the ability to act as if stimuli are related) and (b) highlights that events during learning procedures function in the same way as other cues that are known to control AARR.

Arbitrarily Applicable Relational Responding in a Nutshell

To understand the idea that learning phenomena can be instances of AARR, it is necessary to first explain the concept of AARR. AARR is itself a form of operant behavior, that is, behavior that is a function of its antecedents and consequences (Skinner, 1953). Many instances of operant behavior have as antecedent a single stimulus that signals when the behavior is followed by an outcome. For instance, the presence of a tone might signal that lever pressing will be followed by food. In this case, the tone is the antecedent or discriminative stimulus (Sd), lever pressing is the operant response (R), and food is the consequence or reinforcing stimulus (Sr). If, due to this regularity, lever pressing is more frequent when the tone is present than when the tone is absent, one can say that stimulus control is being exerted (i.e., the tone controls the lever pressing response). In this case, lever pressing would qualify as a non-relational response because it is controlled by an individual stimulus (i.e., the tone).

Operant behavior can, however, also be controlled by a relation between stimuli (see Stewart & McElwee, 2009, for a detailed discussion). Imagine that two tones are presented consecutively and that lever pressing is followed by food only if the duration of the first tone is shorter than that of the second tone. If lever pressing is more frequent when the first tone is shorter than the second tone than when the first tone is longer than the second tone, one can say that lever pressing qualifies as a relational response, that is, an operant behavior that is controlled by a relation between stimuli, that is, a stimulus relation. This would be an example of non-arbitrarily applicable relational responding (NAARR) because the impact of the

relation that functions as an Sd is grounded in the physical properties of the stimuli (i.e., the duration of the tones). [1]

AARR is also operant behavior that is controlled by a relation between stimuli but now the relation that controls responding is not grounded in physical properties. We can illustrate this idea using the example of stimulus equivalence (Sidman, 1971). Consider the symbolic matching-to-sample procedure that is depicted in Figure 1.

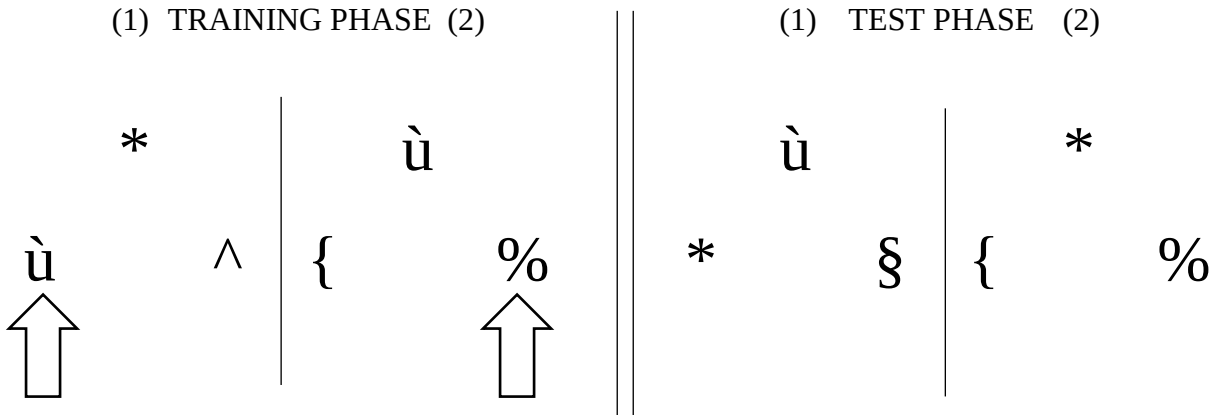

Figure 1. An example of a procedure for studying AARR. The arrows indicate the correct response (taken from De Houwer & Hughes, 2020, Figure 3.9).

During a training phase, participants are reinforced for selecting a particular comparison stimulus in the presence of a particular sample stimulus (e.g., select ù in the presence of * and select % in the presence of ù). Given appropriate controls (e.g., counterbalancing the side on which comparison stimuli are presented), it can be established that these choices are operant behaviors that are under the control not of one stimulus but two stimuli. Which comparison stimulus is the correct option for which sample stimulus is, however, determined by the

[1] As Stuart and McElwee (2009) correctly pointed out, the conclusion that relational responding has occurred, requires evidence that the effect of training generalizes to novel stimuli that are related in similar ways (e.g., short or long tones of a different absolute duration, or lights that are turned on for a short or long duration). If the effect of training does not generalize, it is possible that behavior is controlled not by the relation between stimuli (e.g., longer, shorter) but by individual stimuli (e.g., if tone of 3 sec first, then press lever) or combinations of stimuli (e.g., if tone of 3 sec followed by tone of 5 sec, press lever).

researcher in arbitrary manner, that is, irrespective of the physical properties of the stimuli. Hence, it is unlikely that responding is based solely on relations between physical properties of the stimuli. In principle, responding during the learning phase could be based solely on the direct reinforcement history of the (combinations of) stimuli presented during the learning phase (e.g., when ù and * are on the screen, pick ù) but this cannot account for the choices that people make during a subsequent test phase that involves stimulus displays that they never encountered before. For instance, during the test trials depicted in Figure 1, participants will select above chance stimulus * in the presence of stimulus ù, as well as stimulus % in the presence of stimulus *.

Such a pattern of choices is often referred to as stimulus equivalence. Although it has been studied primarily as a phenomenon it is own right (e.g., Sidman, 1994; Zentall et al., 2014), proponents of RFT (Hayes et al., 2001) have argued that it is one of several types of AARR. More specifically, equivalence responding can be thought of as a type of AARR that involves acting as if stimuli are similar. Like many other behaviors (e.g., preparing a meal), the behavior of "acting as if stimuli are similar" is a behavioral class that encompasses many different behaviors. For instance, acting as if ù, *, and % are similar involves, amongst other things, (a) selecting ù in the presence of *, (b) selecting * in the presence of ù, and (c) selecting % in the presence of *. Stimulus equivalence also involves a transfer of function. For instance, if * has the function of predicting a shock, then acting as if ù is equivalent to * involves acting as if ù also predicts a shock. RFT highlights that people can also act as if stimuli are related in other ways. For instance, they can act as if * is opposite to ù and ù is opposite to %. This would involve several behaviors such as selecting % in the presence of *

or – if * predicts a shock - an increase in fear after the presentation of % but decrease in fear after the presentation of ù (e.g., Dymond & Barnes, 1996). [2]

Because acting as if stimuli are related cannot be grounded solely in the physical properties of the related stimuli (i.e., it occurs even when stimuli are matched arbitrarily) or the reinforcement history of those stimuli during the experiment (i.e., it occurs even in new situations), there must be other factors in play.[3] Two factors have been put forward in the literature (e.g., Hayes et al., 2001): (1) an extensive prior learning history that gives rise to the behavioral repertoire of acting as if stimuli are related and (2) contextual cues that control when and how this repertoire is brought to bear in the current situation. Although there is little empirical research on the first factor, it is assumed that the ability to respond as if stimuli are related in a particular way (e.g., are equivalent) probably arises early on in childhood in social contexts in which adults often encourage children to act as if stimuli are related in a particular way (e.g., to point to an actual dog when hearing the word DOG; Barnes-Holmes & Harte, 2022; Hayes et al., 2001; Hayes & Sanford, 2014). Initially, each individual behavior within a behavioral class (e.g., acting as if stimuli are similar) needs to be trained for each set of stimuli (e.g., reinforce pointing to a dog when hearing DOG; reinforce saying DOG when seeing a dog). After training with many exemplars, however, behaviors within the class can

[2] As Catania (2013, 117–127; also see De Houwer & Hughes, 2020, p. 120) elegantly explains, all operant behaviors can be understood as classes of responses rather than as individual responses. Lever pressing, for instance, can be performed in many different ways (e.g., with different limbs). Descriptively, the class of lever pressing responses is delineated by a researcher-defined criterion (also referred to as the unit of behavior) such as the distance by which a lever is moved downwards. Descriptively, also AARR involves response classes that include many individual responses that meet a researcher-defined criterion, be it that this criterion is defined more abstractly (e.g., acting as if stimuli are similar; acting as if stimuli are opposite). For a more detailed discussion about the operant nature of AARR, please consult D. Barnes-Holmes and Y. Barnes-Holmes (2000).

[3] This argument is still under debate. Some have claimed that stimulus equivalence and related phenomena can arise merely as the result of the reinforcement history of the stimuli during the experimental procedure (see Zentall et al., 2014, for a review). It is difficult to see, however, how this could account for the full complexity of stimulus equivalence, for other types of AARR (e.g., responding as if stimuli are opposite), or for the fact that human adults can flexibly switch between different types of AARR (e.g., Hughes & Barnes-Holmes, 2014).

emerge in new situations without training each of these behaviors (e.g., pointing to a dog when hearing the French word CHIEN; saying CHIEN when seeing a dog).

Whether and how the behavior of AARR is applied in a specific situation is determined by contextual cues. For instance, the mere fact of being reinforced for selecting one stimulus (e.g., ù) in the presence of another stimulus (e.g., *) could function as a contextual cue for responding to these stimuli as if they are equivalent (e.g., also selecting * in the presence of ù even if this has never been trained before; De Houwer & Hughes, 2020). Likewise, research suggests that the mere fact of presenting two stimuli together in space and time can function as a cue for responding to these stimuli as being equivalent (Leader et al., 1996). It is assumed that these events can function as contextual cues for equivalence because of a long history of past learning experiences (e.g., past events in which the cues were present when similar stimuli had to be picked or were presented together). Nevertheless, the impact of those contextual cues can itself be moderated by other contextual cues. For instance, when the selection of ù in the presence of * is reinforced in the context of the word OPPOSITE, people will afterwards respond as if ù and * are opposite (e.g., Steele & Hayes, 1992). Hence, under these conditions, reinforcement of a choice functions as a contextual cue for acting as if two stimuli are opposite. In sum, AARR is relational responding that is grounded in a long learning history that influences current performance in a highly context dependent manner.

Learning Phenomena as AARR

The claim that learning phenomena in humans are instances of AARR has been made most explicitly for an effect known as evaluative conditioning (Hughes et al., 2016a; also see De Houwer & Hughes, 2016, 2017, 2020). As we noted at the start of this paper, evaluative conditioning refers to a change in liking of a conditioned stimulus (CS; e.g., a neutral brand name) that results from pairing that stimulus with a liked or disliked unconditioned stimulus (US; e.g., a picture of smiling people; see Moran et al., in press, for a review). In line with the

suggestion by Leader et al. (1996), Hughes et al. argued that the pairing of a neutral CS and a positive or negative US could function as a cue for the equivalence of the CS and the US. Participants would therefore respond as if the CS and US are equivalent, which includes similar evaluative responses to the CS and US (e.g., liking the CS when the US is also liked).

This analysis not only illustrates what it means to think of a learning phenomenon as AARR but also allows us to explain why there is no contradiction in saying that a behavioral phenomenon can be at the same time a learning effect and an instance of relational behavior. On the one hand, evaluative conditioning can be thought of as a learning effect because evaluative responding to the CS changes as the result of the CS-US pairings (De Houwer, 2007). From this perspective, the behavior of interest is the evaluative response to the CS and the change in this behavior is said to be function of the CS-US pairings. On the other hand, evaluative conditioning (i.e., the change in liking due to stimulus pairings) can be thought of as an instance of relational behavior because it is part of acting as if the CS and US are equivalent in response to the CS-US pairings. From this perspective, the behavior of interest is not the response to the CS but the response to the CS-US pairings: because of the CS-US pairings, participants perform the previously acquired behavior of "acting as if two stimuli are equivalent" for the currently paired CS and US stimuli. As part of this response to the CS-US pairings, the evaluative response to the CS changes, but the response "acting as if two stimuli are equivalent" does not change, it is merely applied to the current CS and US. [4]

There is already some evidence supporting the idea that evaluative conditioning effects are instances of AARR. Hughes et al. (2019), for instance, showed that evaluative conditioning, just like AARR, is context dependent. In addition to presenting CS-US pairs,

[4] Note, however, that emitting a behavior in a new situation can change the *future* likelihood of that behavior. As is the case with any operant behavior, each time that an arbitrarily applicable response such as acting as if two stimuli are equivalent is emitted, this constitutes a new learning episode in that the outcome of emitting the behavior determines the future likelihood of emitting that behavior.

they presented context pairs. For some participants, the context pairs consisted of identical words (e.g., UP – UP) whereas for other participants, they consisted of words with an opposite meaning (e.g., UP – DOWN). Hughes et al. argued that the context pairs should modulate whether or to which the extent CS-US pairings evoke the response of treating the CS and US as equivalent. Whereas context pairs with identical stimuli would confirm that the pairing of stimuli is a cue for equivalence, context pairs with opposite stimuli might reduce the impact of stimulus pairings as a cue for equivalence because the context pairs highlight that (in the context of the experiment) opposite stimuli can also be paired. Evaluative conditioning was indeed stronger when context pairs consisted of identical words then when they consisted of words with an opposite meaning.

The idea that evaluative conditioning effects can be instances of AARR is also in line with the observation that mere instructions about CS-US pairings suffice to induce changes in liking (Hughes et al., 2016a). In fact, from the perspective of AARR, the pairing of stimuli functions in much the same way as the instruction “the CS is equivalent to the US”: both events are cues for responding as if the CS has the same valence as the US. This would explain why the effects of actual pairings are highly similar to the effects of instructions about CS-US pairings (see De Houwer et al., 2020, for a review).

In more recent work, Boddez et al. (2021) suggested that fear conditioning also can be conceived of as an instance of AARR. Fear conditioning refers to the fact that organisms respond fearfully to a CS that reliably precedes an aversive US. Boddez et al. argued that these fearful responses are part of acting as if the CS is similar to other, known predictors of aversive events. More specifically, the fact that the CS reliably precedes the aversive US would function as a cue for acting as if the CS is equivalent to stimuli that predicted aversive events in the past. As a result, people will respond to the CS in the same way as they responded to predictors of aversive events in the past, for instance, by displaying signs of fear.

This perspective sheds new light on the fact that conditioned responses (e.g., freezing in response to CS) can differ drastically from unconditioned responding (e.g., jumping in response to the US). The idea that fear conditioning can also be an instance of AARR is compatible with this divergence because it implies that the CS is not responded to as equivalent to the US (in which case the CS would be responded to in the same way as the US) but as equivalent to other, previously established predictors of aversive events (in which case the CS and US would be responded to differently). Whereas some responses to established predictors of aversive events seem to be hard-wired (e.g., Bolles, 1972), other responses might be learned during the lifetime of the organism. Hence, it should be possible to influence some aspects of conditioned fear responding by influencing the nature of conditioned responding to other, previously established CSs.

Implications

Although the idea that certain learning phenomena qualify as instances of AARR is currently still speculative, we believe that it is worth exploring further because it has some unique implications for learning research. In this section, we focus on implications for (a) empirical and theoretical research on learning, and (b) how learning phenomena in humans relate to other psychological phenomena.

Empirical and Theoretical Implications for Learning Research

If learning phenomena arise because participants deploy a previously established ability for AARR within a current learning procedure, then learning research should have at least two aims: (1) to establish how the ability to AARR is acquired and (2) to uncover the variables that determine how and when this ability is brought to bear. The literature on AARR can provide guidance for these two strands of research. First, there are theories about the nature and timing of the events that are necessary to develop the ability for different types of

AARR (e.g., Barnes-Holmes & Harte, 2022; Hayes et al., 2001; Hayes & Sanford, 2014). As we noted above, the learning history on which AARR relies is very extensive, inherently social, and thus difficult to manipulate in studies. Nevertheless, developmental research in children could shed light on this issue, including research on interventions to influence (e.g., speed up or remedy) the acquisition of the ability for AARR (e.g., Dixon, 2016). Also computational models that simulate the acquisition of the ability for AARR could be helpful in this context.

Second, the literature on AARR contains specific ideas about variables that determine how and when AARR is applied in new situations (see Hughes & Barnes-Holmes, 2016, for a review). If a learning phenomenon qualifies as an instance AARR, then it should be sensitive to the variables that are known to influence AARR, that is, it should have the same functional properties as AARR. This idea already inspired research on evaluative conditioning that tested whether evaluative conditioning, like AARR, is context dependent (e.g., the study of Hughes et al., 2019, that we discussed in the previous section). Future research could extend this to other types of learning (e.g., fear conditioning). It could also test the idea that different aspects of AARR tend to converge. For instance, if CS-US pairings result in a change of liking of the CS because participants act as if the CS and US are equivalent, then those changes in liking should occur together with other aspects of acting as if the CS and US are equivalent (e.g., selecting the CS in the presence of the US within a symbolic matching-to-sample task; see Hughes et al., 2016b).

The proposal that learning phenomena can be instances of AARR inspires not only new strands of research but also new theoretical models. Most existing computational models of learning (e.g., Schmajuk, 2010) are primarily bottom-up models that are determined by events that take place during a learning procedure (e.g., CS-US pairings). This bottom-up approach is also dominant in machine learning research (e.g., Rahwan et al., 2019). Recent

years have seen a stark increase in the use of these approaches in various domains, accompanied by a significant increase in performance, most notably since the introduction of deep learning (i.e., artificial neural networks with multiple hidden layers between the input and output layers). Relative to humans, however, deep learning approaches learn very slowly (i.e., they typically require large sets of labeled training data) and have a limited ability to flexibly generalize problem solutions to novel domains (e.g., Zhang et al., 2018). Our perspective on learning phenomena as instances of relational behavior highlights that humans leverage background knowledge (i.e., knowledge about patterns of relational responding and about events that signal when these patterns should be emitted) to efficiently and flexibly change the way they respond to their environment. Recent research has shown promise in modelling this aspect of human learning both with regard to the development of various techniques to boost efficiency and flexibility by implementing prior knowledge (e.g., Roychodhury et al., 2021) as well as novel insights in how relational information about the environment (states) can be represented and learned from experience (e.g., Doumas et al., 2018, in press).

Despite these developments, we are still far removed from a computational model that captures the flexible nature of AARR. To model (learning phenomena that are instances of) AARR, computational models need to be equipped with information about different patterns of relational responding (e.g., how to respond as if stimuli are equivalent or opposite) as well as some information about contextual cues that signal when and how to deploy this information (e.g., that pairings are a cue for equivalence). Current computational models can encode information about current events (e.g., the fact that a CS co-occurs with or reliably predicts the presence of a US) but they are not equipped with the tools necessary to use this information as cues for relational responding (i.e., information about patterns of relational responding and information about which current events cue which pattern of relational

responding). In sum, the idea that learning phenomena are instances of AARR can provide a source of inspiration for computational modelers to further improve the architecture of their models of those learning phenomena.

Whatever these computational models will look like, they will need to somehow encode information about specific relations (e.g., equivalence, opposition) to model the ability to act as if stimuli are related in a particular manner. In cognitive terms, this means that those models need to encode propositional information. Some learning researchers (e.g., De Houwer, 2009, 2018; Mitchell et al., 2009) have already argued that important learning phenomena are mediated by propositional representations, that is, mental representations that encode how stimuli are related. For instance, evaluative conditioning would arise only after a participant has formed the proposition that the CS co-occurs with the US (De Houwer, 2018, for more details). They contrasted these models with simple association formation models of learning that do not encode relational information but only register covariance, that is, how stimuli covary in the environment (e.g., Rescorla & Wagner, 1972). Because stimuli that are related in different ways might covary in the same way (e.g., a substance in the blood and a disease might covary because the substance causes the disease or because it is an effect of the disease; see Lagnado et al., 2007), simple association formation models cannot capture relational information. Determining how stimuli are related requires not only information about how stimuli covary in the current environment but also other knowledge about the nature of the stimuli and the context in which they occur (e.g., instructions that specify that the substance in the blood is a potential cause of the disease).

From the above, it should be clear that the idea that learning phenomena can be instances of AARR is more compatible with propositional theories of those phenomena than with simple associative accounts. Although the former idea does not require assumptions about mental representations (i.e., it refers only to learning phenomena and AARR as

behavioral effects; see Hughes et al., 2016a), from a cognitive perspective, it makes sense to assume that relational behavior is mediated by relational representations (i.e., propositions; De Houwer et al., 2016). Even though propositional models of learning have been around for some time, there is still novelty and merit in putting forward the idea that some learning phenomena might qualify as AARR. Most importantly, it sidesteps difficult and potentially unproductive debates about the nature of mental representations that mediate learning effects. Because mental representations and operations cannot be observed directly, it is notoriously difficult to reach consensus about their nature. In hindsight, it is therefore perhaps unsurprising that the debate between proponents of association formation models and propositional models of learning did not lead to a consensus (see Boddez et al., 2017, and McLaren et al., 2014, for opposing perspectives). We can, however, sidestep this debate by focusing on whether learning phenomena are instances of AARR. As noted above, this idea does not require assumptions about mental representations. It only specifies that a particular learning phenomenon has the functional properties of AARR (e.g., reliance on prior learning experiences, contextual control, convergence of different components of relational responding). Verifying whether an instance of learning has the functional properties of AARR can feed into the development of cognitive models, but it also has merit on its own in that it allows for an exchange of knowledge about AARR and knowledge about learning.

Implications for How Learning in Humans Relates to Other Psychological Phenomena

The idea that learning phenomena can be instances of AARR not only sets a new agenda for empirical and theoretical research on learning, but also provides a new perspective on how learning phenomena relate to other psychological phenomena. All learning phenomena involve a change in responding to stimuli. In this paper, we advocated the idea that in learning phenomena such as symbolic matching-to-sample learning, evaluative conditioning, and fear conditioning, changes in responding occur because spatio-temporal

regularities (i.e., reinforcing a choice; pairing stimuli) function as contextual cues for relational responding (see De Houwer et al., 2013; De Houwer & Hughes, in press, for a detailed discussion of the core role of spatio-temporal regularities in learning research). However, from AARR research, we know that all kinds of events can function as contextual relational cues. Hence, change in responding can also occur as the result of contextual relational cues other than spatio-temporal regularities. Consider the well-known minimal group effect (e.g., Otten, 2016): merely informing people that an unknown person belongs to the same arbitrary group as a known person results in a change in behavior toward the unknown person. More specifically, people will respond to the unknown person in the same way as they respond to the known person of the same group. Hughes et al. (2020) argued that also this effect can be seen as an instance of AARR in which the sharing of group membership functions as a contextual relational cue for responding to stimuli as if they are equivalent also in other ways. From this perspective, the minimal group effect differs from learning phenomena such as classical conditioning only with regard to the nature of the event that functions as the contextual relational cue: the sharing of group membership versus the pairing of stimuli.

Interestingly, these two types of events have an element in common: both involve similarity between stimuli. Whereas the sharing of group membership involves similarity in terms of group membership, the pairing of stimuli involves similarity in terms of spatio-temporal properties (i.e., the CS and US occur at a similar time and place). This insight led to the proposal that the mere sharing of features can function as a cue for equivalence, regardless of what feature it is that stimuli share (see Hughes et al., 2020; De Houwer & Hughes, 2020, Box 4.1). [5] This Shared Features Principle encompasses many phenomena in psychology (see

[5] Note that the shared features principle can be extended to other relations. For instance, it could be argued that stimuli that are opposite with regard to one feature (e.g., group membership) will be responded to as if they are opposite also with regard to other features. More generally, it could be argued that stimuli that are

Hughes et al., 2020) and clarifies how learning phenomena that are instance of AARR relate to other instances of AARR: they are instances of AARR that involve one specific type of contextual relational cue, namely similarity in terms of spatio-temporal properties.

Until now we have been very careful in saying that *some* learning phenomena *might* be instances AARR. We believe that this level of prudence is appropriate given that this idea is relatively new and is not yet backed up with extensive research. We also realize that the concept of AARR itself is still somewhat controversial (e.g., Zentall et al., 2014; but see Hughes & Barnes-Holmes, 2014). At the same time, it is unlikely that all learning phenomena would qualify as instances of AARR. The most important argument for this position is that full-fledged AARR seems to occur only in verbally able humans. There is no doubt that non-human animals can respond to non-arbitrary relations (e.g., relations grounded in physical properties such as size) but there is little evidence in non-human animals for the flexible deployment of the various types of AARR that verbally able humans display (see Hughes & Barnes-Holmes, 2014, for a discussion). Given that learning phenomena can be instances of AARR only in organisms that are able to show AARR, this already drastically limits the scope of our proposal. Note, however, that even if learning phenomena can be instances of AARR only in verbal humans, our proposal would still have far-reaching implications for learning research. First, all the implications discussed above would still hold for research on learning in verbal humans. Second, it suggests that there might be an important divide in how verbal human beings learn and how other organisms (non-verbal human beings and non-human animals) learn (e.g., Hughes & Barnes-Holmes, 2014). Amongst other things, this would challenge the idea that learning research in non-human animals reveals how (verbal) humans learn. Note, however, that this assumption has been challenged also in the past, for

known to be related in one way (e.g., equivalence, opposition) will also be related in that way with regard to other features.

instance, by the suggestion of Skinner (1966) that learned behavior in humans appears to be rule-governed whereas in other organisms it appears to be contingency-shaped. Because the concept of AARR is a direct descendent of the idea of rule-governed behavior (see Hayes et al., 2001, for a discussion about the relation between the two concepts), Skinner's proposal has very similar implications as the proposal that many learning phenomena in verbal humans are instances of AARR.

In this context, it is interesting to point at one important divergence between, on the one hand, propositional theories of learning and, on the other hand, the idea that learning effects in verbal humans can be instances of AARR (see De Houwer et al., 2016, for a more detailed discussion of this issue). Whereas these ideas are compatible with each other in many ways (see above) they differ in their implications for the relation between learning in verbal versus non-verbal organisms. From the perspective of propositional theories, propositional representations would be necessary for all relational responding, that is, both NAARR and AARR. Given that at least some non-verbal organisms can show NAARR, propositional theories of learning therefore imply that there is no clear divide between how verbal and non-verbal organisms learn: also learning in non-verbal organisms is assumed to be mediated by propositional representations (see De Houwer et al., 2016; Mitchell et al., 2009, pp. 234-235). One way to reconcile propositional theories with the idea that only verbal humans show (learning as an instance of) AARR is to postulate that different kinds of propositional representations or processes underlie NAARR and AARR (De Houwer et al., 2016). What those differences might be, however, is yet to be determined.

Conclusion

In this paper, we clarified and advocated the idea that learning phenomena might sometimes qualify as instances of behavior, more specifically as AARR. Although this idea has until now received little attention from learning researchers, we believe that it potentially

has profound implications for research on learning. We therefore hope that our paper stimulates further discussion and research on the relation between instances of learning and AARR.

References

Barnes-Holmes, D., & Barnes-Holmes, Y. (2000). Explaining complex behavior: Two perspectives on the concept of generalized operant classes. *The Psychological Record, 50*, 251-265.

Barnes-Holmes, D., & Harte, C. (2022). Relational frame theory 20 years on: The Odysseus voyage and beyond. *Journal of the Experimental Analysis of Behavior, 117*, 240-266. https://doi.org/10.1002/jeab.733

Boddez, Y., De Houwer, J., & Beckers, T. (2017). The inferential reasoning theory of causal learning: Towards a multi-process propositional account. In M. Waldmann (Ed.), The *Oxford Handbook of Causal Reasoning* (pp. 1-22). Oxford, UK: Oxford University Press.

Boddez, Y., Finn, M., & De Houwer, J. (2021). The (shared) features of fear: Toward the source of human fear responding. *Current Opinion in Psychology, 41*, 113-117.

Bolles, R. C. (1972). The avoidance learning problem. *Psychology of Learning and Motivation, 6*, 97–145.

Bouton, M. E. (2016). *Learning and behavior: A contemporary synthesis* (2nd Edition). Sinauer Associates.

Catania, A. C. (2013). *Learning* (5th ed.). Sloan Publishing.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10*, 230-241.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*, 1-20.

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin, 13(3*), Article e28046. https://doi.org/10.5964/spb.v13i3.28046

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review, 20*, 631-642. https://doi.org/10.3758/s13423-013-0386-3.

De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and persuasion. *Social Cognition, 34*, 480-494.

De Houwer, J., & Hughes, S. (2017). Environmental regularities as a concept for carving up the realm of learning research: Implications for Relational Frame Theory. *Journal of Contextual Behavioral Science, 6*, 343-346.

De Houwer, J., & Hughes, S. (2020). *The psychology of learning: A functional-cognitive introduction.* The MIT Press.

De Houwer, J., & Hughes, S. (in press). Learning in individual organisms, genes, machines, and groups: A new way of defining and relating learning in different systems. *Perspectives on Psychological Science*.

De Houwer, J., Hughes, S., & Barnes-Holmes, D. (2016). Associative learning as higher-order cognition: Learning in human and nonhuman animals from the perspective of propositional theories and Relational Frame Theory. *Journal of Comparative Psychology, 130*, 215-225.

De Houwer, J., Van Dessel, P., & Moran, T. (2020). Attitudes beyond associations: On the role of propositional representations in stimulus evaluation. *Advances in Experimental Social Psychology, 61*, 127-183.

Dixon, M. R. (2016). *The PEAK relational training system: Transformation module*. Shawnee Scientific Press.

Doumas, L. A., & Martin, A. E. (2018). Learning structured representations from experience. In *Psychology of Learning and Motivation* (Vol. 69, pp. 165-203). Academic Press.

Doumas, L. A., Puebla, G., Martin, A. E., & Hummel, J. E. (in press). A theory of relation learning and cross-domain generalization. *Psychological Review*.

Dymond, S., & Barnes, D. (1996). A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness and opposition. *The Psychological Record, 46*, 271-300.

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

Hayes, S. C., & Sanford, B. T. (2014). Cooperation came first: Evolution and human language and cognition. *Journal of the Experimental Analysis of Behavior, 101*, 112–129. https://doi.org/10.1002/jeab.64

Hughes, S., & Barnes-Holmes, D. (2014). Associative concept learning, stimulus equivalence, and relational frame theory: Working out the similarities and differences between human and non-human behavior. *Journal of the Experimental Analysis of Behavior, 101*, 156–160.

Hughes, S., & Barnes-Holmes, D. (2016). Relational frame theory: The basic account. In S. Hayes, D. Barnes-Holmes, R. Zettle, and T. Biglan (Eds.), *Handbook of Contextual Behavioral Science* (pp. 129–178). Wiley.

Hughes, S., De Houwer, J., & Barnes-Holmes, D. (2016a). The moderating impact of distal regularities on the effect of stimulus pairings: A novel perspective on evaluative conditioning. *Experimental Psychology, 63*, 20-44.

Hughes, S., De Houwer, J., Mattavelli, S., & Hussey, I. (2020). The Shared Features Principle: If two objects share a feature, people assume those objects also share other features. *Journal of Experimental Psychology: General*, *149*, 2264–2288.

Hughes, S., De Houwer, J., & Perugini, M. (2016b). Expanding the boundaries of evaluative learning research: How intersecting regularities shape our likes and dislikes. *Journal of Experimental Psychology: General*, *145*, 731-754.

Hughes, S., Ye, Y., & De Houwer, J. (2019). Evaluative conditioning effects are modulated by the nature of contextual pairings. *Cognition & Emotion, 33*, 871–884.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–72). Oxford University Press.

Leader, G., Barnes, D.,&S meets, P.M.(1996).Establishing equivalence relations using a respondent-type training procedure. *The Psychological Record, 46*(4), 685–706.

McLaren, I. P., Forrest, C., McLaren, R., Jones, F., Aitken, M., & Mackintosh, N. (2014). Associations and propositions: The case for a dual-process account of learning in humans. *Neurobiology of Learning and Memory, 108*, 185–195.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*, 183-198.

Moran, T., Nudler, Y., Bar-Anan, Y. (in press). Evaluative Conditioning: Past, Present, and Future. *Annual Review of Psychology*.

Otten, S. (2016). The minimal group paradigm and its maximal impact in research on social categorization. *Current Opinion in Psychology, 11*, 85–89.

Rahwan, I., Cebrian, M., Obradovich, N. et al. (2019). Machine behaviour. Nature 568, 477-486. https://doi.org/10.1038/s41586-019-1138-y

Rahwan, I., Cebrian, M., Obradovich, N. et al. (2019). Machine behaviour. *Nature*, *568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2*, 64–99.

Roychowdhury, S., Diligenti, M., & Gori, M. (2021). Regularizing deep networks with prior knowledge: A constraint-based approach. *Knowledge-Based Systems*, *222*, 106989.

Schmajuk, N. A. (2010). *Computational Models of Conditioning*. Cambridge University Press.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research, 14*, 5-13.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA: Authors Cooperative.

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Macmillan.

Skinner, B. F. (1966). An operant analysis of problem solving. In B. Kleinmutz (Ed.), *Problem solving: Research, method and theory* (pp. 225-257). Wiley.

Steele, D., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding.*Journal of the Experimental Analysis of Behavior*, *56*(3), 519-555.

Stewart, I., & McElwee, J. (2009). Relational responding and conditional discrimination procedures: An apparent inconsistency and clarification. *The Behavior Analyst, 32*, 309–317.

Zentall, T. R., Wasserman, E. A., & Urcuioli, P. J. (2014), Associative concept learning in animals. *Journal of the Experimental Analysis of Behavior*, *101*, 130-151. https://doi.org/10.1002/jeab.55

Zhang, C., Vinyals, O., Munos, R., & Bengio, S. (2018). A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*.