

A partially hardcoded computational model of arbitrarily applicable relational responding

Matthias Raemaekers^{*} , Martin Finn , Jan De Houwer 

Ghent University, Belgium

ARTICLE INFO

Keywords:

Arbitrarily applicable relational responding
Relational frame theory
Computational modelling
Reinforcement learning

ABSTRACT

Relational Frame Theory (RFT) has inspired a considerable body of research demonstrating fundamental aspects of arbitrarily applicable relational responding (AARR) purported to be the building blocks of human language and cognition. Current empirical research has certain limitations, however. Computational modelling allows researchers to circumvent certain practical limitations to studying (the development of) complex forms of AARR and force them to scrutinize the theory, yet limited research has sought to develop computational models of AARR. RFT can guide researchers developing computational models of AARR by emphasizing its operant and context-sensitive nature and by clarifying the nature of the learning history that is required for its development. We describe a partially hardcoded computational model of AARR inspired by RFT. By combining computational reinforcement learning with hardcoded relational knowledge, we simulate AARR as observed in a conceptual replication of the seminal Steele and Hayes (1991) study. Our results add to a growing literature modeling relational behavior and illustrate that RFT can be useful for researchers interested in computational modelling. Implications and future research directions are discussed.

Relational frame theory (RFT, Hayes et al., 2001) has been influential in stimulating empirical research on relational responding (Hayes et al., 2021; O'Connor et al., 2017), new approaches in behavior-analytic therapy (i.e., Acceptance and Commitment Therapy or ACT; Hayes, 2004; Hayes et al., 2011) and interventions for relational responding abilities (e.g., Cassidy et al., 2011; Dixon, 2014). Central to RFT is the claim that the human ability for arbitrarily applicable relational responding (AARR) is the cornerstone of human language and cognition. Relational responding refers to the ability to respond to one event in terms of its relation to another (e.g., choosing the larger of two apples at the supermarket). Crucially, these relations need not be defined by formal stimulus properties (e.g., size, color, shape), but can be defined solely by contextual cues (e.g., verbal, social cues) that indicate what stimulus function should be related (e.g., reinforcing, fear, valence, etc.), and in what manner (e.g., sameness, distinction, comparison, etc.). For instance, humans familiar with the euro currency will act as if a small coin (1 euro) is more valuable than a larger coin (0.5 euro) despite their relation in terms of physical size. We refer to such contextually defined relational responses as AARR.

Research has demonstrated how reinforcement contingencies can shape up patterns of AARR and bring them under contextual control (e.

g., Delabie et al., 2022; Steele & Hayes, 1991). Research has also provided evidence that stimuli can acquire a wide array of behavioral functions (self-discrimination: Dymond & Barnes, 1994, 1995, 1996; respondent eliciting: Dougher et al., 1994; sexual arousal: Roche & Barnes, 1997; discriminative: Roche et al., 2000; consequential: Whelan & Barnes-Holmes, 2004; avoidance: Dymond et al., 2008; meaning, Perez et al., 2023) by being arbitrarily related to other stimuli, in accordance with many types of relations (e.g., sameness, distinction and opposition: Dymond & Barnes, 1994, 1995, 1996; comparison: Barnes-Holmes et al., 2004; Whelan & Barnes-Holmes, 2006; deictic: McHugh et al., 2004). Furthermore, research demonstrated that stimulus functions can be transformed by relating relations (Barnes et al., 1997; Lipkens & Hayes, 2009; Stewart et al., 2001, 2002, 2004), and by relating entire relational networks (O'Hora et al., 2004; O'Hora et al., 2014; Perez et al., 2023).

Despite this impressive body of empirical demonstrations, there remains a gap between relational behavior observed in the lab and the real-life symbolic behavior that is the ultimate target of RFT (e.g., language, metaphor, complex problem-solving, reasoning; McIlvane, 2003). We believe that alternative approaches beyond behavioral laboratory studies could facilitate empirical research on this complex

^{*} Corresponding author. Dunantlaan 2, Gent, 9000, Belgium.

E-mail address: Matthias.raemaekers@ugent.be (M. Raemaekers).

behavior. Furthermore, there is relatively little longitudinal research in individual agents on how and when human children start displaying the ability to AARR (but see Lipkens et al., 1993; Luciano et al., 2007; Cassidy et al., 2011). If we cannot analyze (the development of) complex relational behavior in the lab, it is difficult to devise effective procedures to influence it outside of the lab. Admittedly, setting up the conditions and learning history for AARR to develop in the lab is difficult (Dougher & Markham, 1994; Lyddy & Barnes-Holmes, 2007; Lipkens & Hayes, 2009), and the learning history itself is rarely described in detail (O'Hora et al., 2004; Lyddy et al., 2001). RFT-researchers, aware of these limitations, have chosen a piecemeal approach, conducting systematic inductive experimental analyses of each element of the RFT-account of human relational responding (Lipkens & Hayes, 2009). However, an assumption underlying this approach is that knowledge of the determinants of AARR within developmental segments will yield a complete picture of the developmental learning trajectory of AARR. This may not be the case. That is, the picture of a learning history resulting from such a strategy may differ in important ways from that obtained by linear studies of the developmental trajectory. Validating the strategy requires combing the knowledge about these different segments into an (artificially constructed) learning history, and testing whether that history is sufficient to produce AARR as observed in human adults.

Development of computational models of AARR can help researchers overcome some of these barriers. Constructing and simulating a learning history *in silico* is more time- and cost-efficient than *in vivo* studies and allows for direct control over complex learning histories (Lyddy & Barnes-Holmes, 2007; Carrillo & Betancort, 2024). Good computational models of AARR would allow researchers to simulate the outcomes of behavioral experiments or interventions without having to set up complex experimental procedures or longitudinal designs. Moreover, computational research requires scrutiny of the theory (e.g., specifying the learning history required for a system to learn how to AARR) and can provide formal tests of its predictions. We realize that it is not a given that studying artificial agents will ultimately be helpful for predicting and influencing human behavior, nor can we be certain that the complex learning histories constructed for the models are identical to those that afford the development of AARRing in humans. Modelers will need to go beyond what is currently described by RFT literature. However, given the current limitations of empirical research and the increasing availability of modelling tools and literature, we believe the prospect of new knowledge to be worth the effort. In the next section, we describe the core principles of RFT that we argue a computational model of AARR should adhere to. We then review relevant research on computational modeling of AARR, before proposing a novel computational reinforcement learning approach to modelling AARR that we then use to simulate AARR in a conceptual replication of Steele and Hayes (1991).

1. RFT principles for a computational model of AARR

Relational responding is a generalized operant that develops throughout a long and rich history of reinforcement (Hayes et al., 2001). RFT strongly emphasizes that this learning history must be sufficiently rich and appropriately structured. The development of AARR involves the abstraction of generalized patterns of relational responding away from the irrelevant stimulus properties involved in the various relational responses observed across a learning history (i.e., learning what it 'means' for things to be same, different, opposite or in some other way related, in a general, abstract sense). A necessary condition for this abstraction is multiple exemplar training (MET), that is, being reinforced for responding relationally to different stimuli (e.g., interactions involving responding to many different word-referent pairs in early childhood). A second requirement is that these interactions occur in the presence of contextual cues that discriminate which type of relational responding will be reinforced (e.g., the word 'is' in "*that is a cat*" or "*what is that?*"). During MET, patterns of relational responding that are repeatedly reinforced in the presence of these cues can be abstracted and

brought under their control and can then be flexibly applied to novel stimuli or situations when those contextual cues are present. Computational models of AARR must therefore be sensitive to the learning context. Finally, the learning history must be structured in the sense that more complex instances of AARR are predicated on more rudimentary ones (Lipkens et al., 1993). For instance, responding to analogies or relational networks requires the ability to recognize and respond to the relations involved in them (Barnes-Holmes et al., 2020).

When patterns of AARR have been abstracted and are under the control of contextual cues, they can in principle be applied to any novel stimuli when appropriate cues are provided. For instance, people visiting a new country who are told that a currency of small purple coins is *more valuable* than your own can respond accordingly (e.g., pick a purple coin over a dollar) and can derive novel relations (e.g., pick purple coins over another currency that is less valuable than dollars) despite the coins being small and cartoonish. It is worth noting that the latter can be conceived of as a separate learning process, that is, learning how known relations (i.e., already abstracted from a prior MET history) may hold between novel stimuli, and deriving novel relations from that information. This learning process plays out on a smaller timescale and relies on the longer-term development of the general ability to respond relationally (abstraction and contextual control). Such a short-term learning process is what is typically assessed in behavioral experiments. Steele and Hayes (1991) pioneered the procedure that is now predominantly used in such experiments and inspired many empirical demonstrations of the arbitrary, flexible and generative nature of AARR (Hayes et al., 2021).

The procedure introduced by Steele and Hayes (1991) consists of three phases. Phase 1 (referred to as pretraining) typically involves MET of non-arbitrarily applicable relational responding (i.e., stimulus relations are defined by physical properties, Pre-training, Fig. 1) and serves to establish novel contextual cues (i.e., by selectively reinforcing specific relational responses, e.g., select the identical stimulus, in the presence of specific cues). The newly established contextual cues are then used to train baseline relations between arbitrary stimuli (Arbitrary Training, Fig. 1). In Phase 2 (referred to as training) participants are reinforced for selecting one stimulus out of a set of comparison stimuli in the presence of a sample stimulus and a contextual cue, but crucially, no meaningful non-arbitrary relationships exist between the sample and comparison stimuli. Phase 3 (derived relational responding test) involves testing (i.e., without feedback) participants' performance producing untrained stimulus relations, which can be derived from the relations trained in Phase 2 (Fig. 1, Test). The procedure as adapted in this research is described in detail in the Method section.

Arguably, the fact that humans can succeed at such tasks is largely dependent upon a long prior history of reinforcement for relational responding (i.e., a pre-experimental learning history). The (adolescent) participants in the Steele and Hayes (1991) study presumably had prior experience that facilitated abstraction of the relations involved in the (pretraining) task (i.e., what it means for two stimuli to be the same, different or opposite, regardless of the particular stimulus properties that are related in this way), as well as knowledge of contextual cues that signal those relations. Earlier, we described how we can conceptually separate this long-term process of abstraction from the short-term application of acquired, abstract relational knowledge in novel situations (i.e., what is typically referred to as the experimental learning history). Rather than modelling the process of relational abstraction (Hayes et al., 2001, Chapter 2; Penn et al., 2008), it is likely that the contextual cue pretraining phase serves to establish novel contextual control (i.e., novel contextual cues) over the relations that participants already have knowledge of (e.g., by matching the known patterns or relations to the cues encountered in the task; a trivial task compared to developing an abstract understanding of relations). This process of learning about novel contextual cues based on non-arbitrary relationships between stimuli can also be studied separately from learning about arbitrarily applied relations between stimuli as a function of known

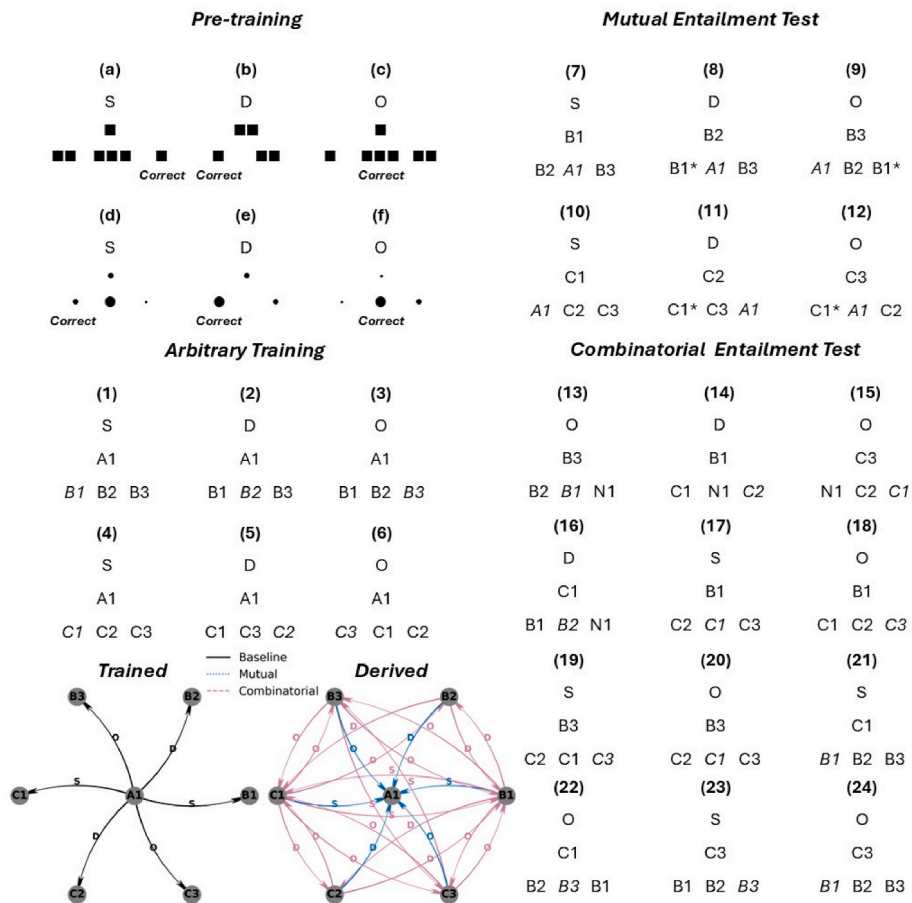


Fig. 1. Steele and Hayes (1991) Procedure as adapted for model simulations.

Note. S, D and O represent relational contextual cues for sameness, difference and opposition, respectively. Correct responses in arbitrary training and testing are printed in italic. Trials 8, 9, 11 and 12 unintentionally (i.e., due to a programming error) presented a second correct comparison stimulus (via combinatorial entailment of baseline relations), indicated with an asterisk.

contextual cues in the environment.

Researchers modelling AARR can target each learning process separately and each may provide unique knowledge and predictions to be tested in future behavioral research. A model of the short-term application of relational knowledge would allow researchers to test what kind of knowledge is required for flexible and generative AARR, make predictions about the relative effectivity of procedures to establish novel behavior, or fit and compare different model versions (e.g., with limited knowledge, biases or different parameters) to model AARR in different groups (e.g., clinical or developing populations). Modelling the establishment of novel contextual cues can help researchers better understand how abstract representations of relations can be learned (e.g., what does the agent need to represent, how much MET is required) and how this may or may not be different for different kinds of cues (e.g., cues for different relations, functional versus relational contextual cues, or novel versus known relations). Furthermore, knowledge gained from modelling the short-term learning processes may also have implications for modelling the long-term learning process (e.g., what kind of experiences are required to produce the knowledge that is used in the short-term process). Finally, modelling the long-term process of abstraction would provide information about the establishment of contextual cues and would allow researchers to learn about the necessary conditions for AARR to emerge, model individual differences in the development of AARR abilities, or simulate the outcomes of experiments or interventions (Binz & Schulz, 2023; e.g., training relational abilities: PEAK, Dixon et al., 2014; or SMART, Cassidy et al., 2011).

2. Past research on computationally modelling AARR

Relatively little research has tried to develop computational models of AARR, contrasting with the increasing use of computational modelling in psychological research (Ninness et al., 2018; Wilson & Collins, 2019) and rapid advances in artificial intelligence research (e.g., OpenAI, 2024). We provide a brief overview of the most relevant literature that inspired the current work, but given the scope of the current paper we cannot go into detail. Readers interested in these details are encouraged to consult the review by Tovar et al. (2023) or the referenced publications. Barnes and Hampson (1993) developed a neural network¹ (NN) named RELNET to model the AARRing described in the seminal Steele and Hayes (1991) match-to-sample (MTS) procedure and the transfer of sequence functions in equivalence classes (Cullinan et al., 1994; Lyddy & Barnes-Holmes, 2007; Lyddy et al., 2001). The model simulated context-dependent AARR as observed by Steele and Hayes (1991) but was criticized for having too much information about the task available to it (Tovar and Chávez, 2012). RELNET did inspire other researchers to develop computational models of stimulus equivalence (see Tovar et al., 2023). Using various approaches, researchers have computationally modeled the formation of equivalence classes in

¹ Neural networks are a type of machine learning algorithm composed of (often many) layers of interconnected artificial neurons. The systems learn to produce the correct output for a given input by gradually adapting the weights connecting the input and output neurons (and layers in between) as a function of a feedback signal (provided by the modeler or the system itself).

humans and typical moderators thereof. That is, NNs simulated the formation of equivalence classes (Lyddy & Barnes-Holmes, 2007; Vernucio & Debert, 2016; Ninness & Ninness, 2020) and showed derived symmetry and transitivity of equivalence relations (Tovar and Chávez, 2012). They also demonstrated similar effects of training protocol (e.g., Lyddy & Barnes-Holmes, 2007; Carrillo & Betancort, 2023, 2024) and network size (e.g., Mofrad et al., 2020) as typically found in humans, which hints at their potential for predicting human behavior in such procedures (Ninness & Ninness, 2020).

Most approaches to modelling AARR have used classical artificial NNs (but see Carrillo & Betancort, 2023, 2024, for an assessment of stimulus equivalence in transformer-based models), which are notoriously black boxes, potentially limiting what researchers could learn about the development of AARR from studying these models. Alternative approaches are more easily interpretable. Mofrad et al. (2020)'s approach, called equivalence projective simulation,² simulated stimulus equivalence in MTS procedures by combining computational reinforcement learning³ (RL, Sutton & Barto, 1998) principles with an episodic memory. This line of research was extended by O'Sullivan et al., (2025) to model relational density theory (Belisle & Dixon, 2020), a recent theoretical development adjacent to RFT, but this work is currently limited to equivalence relations. Other approaches have shown promise modelling AARR beyond stimulus equivalence, like Duran et al., (2026) who used graph neural networks to model core properties of AARR. Finally, the Non-Axiomatic Reasoning System⁴ (NARS, Wang, 2006) takes a fundamentally different, but equally promising approach. NARS uses a non-axiomatic logic to reason about knowledge acquired from experience, and has so far demonstrated human-like generalized identity matching (Johansson et al., 2022), stimulus equivalence (Johansson et al., 2024; Johansson & Lofthouse, 2023) and generalized derived relational responding to equivalence and opposition relations (Johansson et al., 2025). We encourage this promising research, but note that existing models: i) are limited to only a few relations (primarily stimulus equivalence) whereas AARR can involve many relations (e.g., opposition, comparison, hierarchy, analogy), and ii) have been used in a limited number of task settings (e.g., go/no-go procedures: Vernucio & Debert, 2016; compound stimuli: Tovar and Torres-Chávez, 2012; MTS: Ninness & Ninness, 2020; Mofrad et al., 2020; Johansson et al., 2025). More research in this direction is needed. We therefore decided to adopt a new approach to modelling AARR, which we introduce in the next section.

3. The current research: A computational reinforcement learning model of AARR

To model adult humans that have already acquired the ability to AARR flexibly in novel situations, we make two key assumptions. Earlier, we argued that for AARR to be applied flexibly and generatively in novel situations, the artificial agent must have knowledge of contextual cues (i.e., understanding how a cue discriminates the relation to respond to, and the function to be transformed) and of patterns of relation derivation (i.e., the relations that are mutually and

² This research leverages the adaptivity of reinforcement learning (see Footnote 3) algorithms in an episodic memory that stores equivalence relations as graphs to simulate the formation of equivalence classes in MTS.

³ RL algorithms are a branch of machine learning in which an artificial agent learns by interacting with its environment and changing its behavior as a function of reinforcement it receives for particular interactions. The agent gradually adapts its behavior as a function of the predicted reinforcement, with the goal of maximizing reward.

⁴ NARS is a general-purpose intelligent reasoning system designed to adapt to its environment under insufficient knowledge and resources. It uses so-called non-axiomatic logic to reason about events and was extended to also reason about relations.

combinatorially entailed given a particular set of source relations). We assume that this knowledge is the end result of the long-term learning process (abstraction from MET) described above. Second, the model should have context-sensitive learning mechanisms that allow it to use reinforcement as a basis for learning and deriving relations between novel stimuli. While RFT is a functional-analytic theory and thus does not concern itself with hypotheses about the representational (algorithmic or neural) nature of this knowledge, one cannot escape taking on a (cognitive-) mechanistic perspective when trying to construct a computational model (i.e., the algorithm is a mechanism that mediates the functional relations described by RFT). We do, however, wish to make clear we have no explicit hypotheses about the nature of these mechanisms in humans (e.g., whether they correspond to particular cognitive processes or how they may be implemented in human brains). Instead, we take a pragmatic approach with the aim of modelling AARR in humans in a way that is consistent with RFT's description of it. As pointed out earlier, this approach has potential merit in that it provides a plausibility test of assumptions about how the ability to AARR comes about (i.e., known relational contextual cues that guide the application of acquired patterns of relational responding) and requires us to make those assumptions more explicit. In the method section, we describe in more detail how our approach extends a classical RL algorithm with relational knowledge in a way that is inspired by RFT. In that section, we also describe a conceptual replication of the Steele and Hayes (1991) study that involved using culturally established contextual cues (i.e., the words 'same', 'different' and 'opposite') to establish a network of arbitrarily applied sameness, difference and opposition relations and test for derived responding (i.e., Arbitrary Training and Mutual and Combinatorial Test in Fig. 1). We then describe both the human data and the simulation of their behavior using our RL algorithm.

4. Method

4.1. Participants and design

Seventy-two participants (43 female, 27 male, two non-binary, mean age 35.76 years, $SD = 14.07$ years) were recruited on the Prolific online research platform.⁵ A priori power analysis indicated that a sample of 72 provides 0.8 power to detect a small effect size (i.e., Cohen's $d = 0.3$) at a significance level of $\alpha = .05$ using a one-sample t -test (i.e., compare participant test performance to chance-level). All participants were required to be at least eighteen years old, be native Dutch speakers and to not have participated in prior experiments of our lab. All participants provided informed consent before participating in the experiment. The study was approved by an Ethical Committee and was carried out in line with the declaration of Helsinki.

4.2. Materials

Software. The experiment was programmed using the labjs⁶ online study builder (Henniger et al., 2020). Scripts to run analyses were programmed in R (version, 3.6; R Core Team, 2019). All scripts to run the experiment and analyses were preregistered and are available on the Open Science Framework.⁷

Stimuli. We used the three English words *same*, *opposite* and *different* as contextual cues. Nine other randomly selected geometric shapes (taken from Steele & Hayes, 1991; see Supplementary Materials) were used to train an arbitrary relational network. No readily identifiable relationships existed between the geometric figures and they were randomly assigned to serve as the different stimuli in the network (i.e. labeled A, B1, B2, B3, C1, C2, C3, and N1; counterbalanced between

⁵ <https://www.prolific.com/academic-researchers>.

⁶ <https://labjs.felixhenniger.com/>.

⁷ https://osf.io/8xbnj/?view_only=50bfd49da669403aa307830d3757cb44.

participants).

4.3. Procedure

Arbitrary MTS-training. After providing informed consent and demographic information, participants in Experiment 1a started with the arbitrary training phase. The arbitrary MTS phase served to train a network of sameness, difference and opposition relations between seven randomly selected geometric figures (no readily discernible relations between them), using existing (i.e., the words ‘same’, ‘different’ or ‘opposite’, in Experiment 1a) or newly established cues (the nonwords established as relational cues in the pretraining phase of Experiment 1b). Responding in accordance with these (arbitrarily applied) relations was shaped up by selectively reinforcing the matching of a particular comparison stimulus (e.g., B2) to a sample stimulus (e.g., A) in the presence of a particular contextual cue. The different trials of this phase are illustrated in Fig. 1 (Trials 1 through 6). Participants completed six blocks of 36 trials, on which a contextual cue was presented at the top of the screen, a sample stimulus in the center of the screen, and three comparison stimuli at the bottom of the screen. Participants selected a comparison stimulus using the ‘d’- (for the left comparison), ‘g’-(middle) and ‘j’-keys (right) on the keyboard, and were presented feedback (i.e., “Correct!” or “Wrong!”, 1500ms, in green and red, respectively), before the next trial started (1000ms ITI). In between blocks, participants were told how many correct responses they made in the previous block and instructed that they could take a short break before continuing. The number of trials probing each relation was balanced within blocks (i.e., six trials each), as were the locations of the reinforced comparison stimulus (i.e., left, middle or right).

Derived relational responding test. In the final phase, we tested responding in accordance with relations that could be derived from the previously trained (arbitrarily applied) relations. If we assume some abstract understanding of relations, participants were expected to derive novel (mutually and combinatorially entailed) relations between the arbitrary stimuli. Participants were instructed that they would perform a task similar to the previous one, wherein they would no longer be provided feedback on their choices, but that their goal was still to respond as accurately as possible. Trials were presented in the same way as in the arbitrary training phase described above, but no feedback on was presented after participants’ responses. In two blocks of 36 trials (1000ms ITI), we tested the baseline relations trained in the previous phase (Trials 1 through 6 in Fig. 1), and their mutually entailed relations (Trials 7 to 12 in Fig. 1). For instance, if selecting B1 given sample A and cue SAME was reinforced in training, we tested whether participants would select A given sample B1 and cue SAME. Each relation (six trained and six derived) was tested six times (i.e., a total of 72 trials). In a third block (twelve trials), we also tested combinatorially entailed relations (Trials 13 through 24 in Fig. 1). For example, if both selecting B1 and selecting

C1 given sample A and cue SAME, we tested whether participants would select C1 given sample B1 (or vice versa) in the presence of a cue for SAME. Note that on some of those trials, a novel stimulus (N1) was introduced to ensure three comparison stimuli could be presented, of which only one could be considered the correct response.⁸

4.4. Computational reinforcement learning algorithm with relational knowledge

Hardcoded Relational Knowledge. As noted above, typical adult participants in AARR experiments bring with them relational knowledge that their performance in such tasks relies upon. Adult humans are assumed to know the ‘meaning’ of Crels: the fact that they indicate a relation between events (and possibly what particular stimulus function is related). Their learning history should have also allowed them to acquire an abstract understanding of various relations (i.e., what it means for events to be the same as, different from, or opposite to one another) and what other knowledge can be derived from that (i.e., if one knows A is related to B, then they also know B is related to A in some way; and if A is also related to C, then one knows B and C are also related). This relational knowledge is represented in the model as so-

Mutual Entailment.		Combinatorial Entailment.			
Source Relation	Entailed Relation	Relation 1			
Same	Same		S	D	O
Different	Different	Relation 2			
Opposite	Opposite	Same (S)	S	D	O
		Different (D)	D	/	/
		Opposite (O)	O	/	S

Fig. 2. Derivation tables for mutual and combinatorial entailment of sameness, difference and opposition relations.

Note. The left table provides source relations (e.g., A is the same as B) and the relations that they mutually entail (e.g., B is the same as A). The right table shows how two source relations, Relation 1 (e.g., A same as B) and Relation 2 (e.g., A is opposite to C), together afford derivation of new relations (the other cells, e.g., B opposite to C). A forward slash indicates that no clear derivation is possible. Note that technically, the order of relata in combinatorially entailed relations matters (referred to as different training protocols: linear $\bar{A} B$ and $B \sim C$; one-to-many $\bar{A} B$ and $\bar{A} C$; or many-to-one $\bar{A} B$ and $C \sim B$). Different protocols afford different derivations, but only when derivation involves non-symmetric relations (mutual entailment results in a different relation, e.g., if A is more than B, B is less than A). All relations in the current procedure are symmetrical, so we only included one table. Expanded versions of the derivation tables available for consultation on the OSF (https://osf.io/bjqxq/?view_only=5ab9e17676a3436d9d7878d0b7e1340e).

⁸ Originally, a final block of twelve trials also assessed combinatorially entailed relations, but also introduced another novel stimulus (N2) as a comparison, about which participants should be able to via exclusion, if they successfully learned the baseline relations. Three consecutive trials might present C1 as the sample stimulus, and B1, B2 and N2 as comparison stimuli. On the first, the cue SAME is presented (so participants should select B1) and on the second, the cue for DIFFERENT is presented (so participants should select B2). On the third trial, the cue for OPPOSITE is presented, but participants who successfully learned the baseline relations should know that neither B1 or B2 were related as opposite to C1, and are therefore expected to select the novel stimulus N2 (i.e., reasoning by exclusion). On further trials then, N2 is presented as the sample stimulus, and newly derived relations (e.g., N2 is the same as C3) are tested. However, due to a programming error, the trial order in this block was randomized, making it impossible for participants to learn as we intended them to. We choose not to report this phase or the experiment, or the data, here, but they are available on the OSF (https://osf.io/8xbnj/?view_only=50bfd49da669403aa307830d3757cb44).

called derivation tables (illustrated in Fig. 2). These tables provide the reader (or in this case, the artificial agent) with the relations that can be derived from a given (pair of) source relations (see also Allen, 1983, for an application of the same idea to temporal relations between events). For mutual entailment (i.e., if A is related to B, B is related to A), the table consists of one row per relation, and two columns, one for the source relation and one for the derived relation. For combinatorial entailment, columns represent different relations between A and B and rows represent different relations between B and C. Each cell represents the relation that can be derived from that combination of relations (or if no relation can be derived). The next section includes examples of how the model uses the derivation tables to derive novel relational responses.

Reinforcement Learning and Relational Learning. To model AARRing in adults, we use a RL approach. RL is suitable for modelling AARR in that the agent learns directly from (reinforced) interactions with its environment (see Sutton & Barto, 1998). Compared to NNs, RL models are typically easier to interpret, and there is empirical evidence that linked RL models to human learning and neurobiology (e.g., Botvinick et al., 2020; Niv, 2009; O'Doherty et al., 2015; Subramanian et al., 2022).

Value-learning. The model uses a temporal difference learning algorithm akin to Q-learning (Watkins & Dayan, 1992). The agent tracks estimates of two types of values: (1) stimulus values, $V(c_b, s_t)$, representing the expected reinforcement associated with a stimulus s , presented in context c on trial t ; and (2) stimulus-action values, $Q(c_b, s_t, a_t)$, or the reinforcement expected for selecting a comparison stimulus a in the presence of a sample stimulus s and contextual cue c on trial t .

These contextualized value-estimates are updated incrementally as a function of the reward R received on trial t , using a temporal-difference learning delta rule:

$$V_{t+1}(c_t, s_t) \leftarrow V_t(c_t, s_t) + \alpha \cdot \delta_t^V \quad \alpha \in [0, 1] \quad (1a)$$

$$\delta_t^V = R_t - V_t(c_t, s_t) \quad (1b)$$

$$Q_{t+1}(c_t, s_t, a_t) \leftarrow Q_t(c_t, s_t, a_t) + \alpha \cdot \delta_t^Q \quad \alpha \in [0, 1] \quad (2a)$$

$$\delta_t^Q = R_t - Q_t(c_t, s_t, a_t) \quad (2b)$$

α is a learning rate parameter that scales the value-updates (smaller incremental updates for lower learning rates).

Comparison Selection. On any given trial (training or test), the agent stochastically selects a comparison stimulus by comparing estimated values of the available response options and any derived response values. A choice is selected by means of a softmax function (Equation (3)) that returns the weighted probability of selecting comparison stimulus a_t given sample stimulus s_t and cue c_t :

$$P(a_t | c_t, s_t) = \frac{e^{Q_t(c_t, s_t, a_t) / \tau}}{\sum_i^n e^{Q_t(c_t, s_t, a_i) / \tau}} \quad \tau \in [0, 5] \quad (3)$$

τ is a temperature parameter: higher values produce more exploratory (less deterministic) responding.

Stimulus Relations and Relation Derivation. We extended typical RL algorithms with hardcoded knowledge of contextual cues and relational derivation (represented in derivation tables) and with a mechanism to track the relational similarity of stimuli, to support derived relational responding (i.e., mutual and combinatorial entailment). The agent computes a relatedness score, $rel_t(c_b, s_b, a_b)$, based on the similarity of value-estimates for the stimuli in a given context. The relatedness between a sample stimulus s_t and comparison stimulus a_t in context c_t is defined as:

$$rel_t(c_t, s_t, a_t) = e^{-|V_t(c_t, s_t) - V_t(c_t, a_t)|} \quad (4)$$

This function maps the absolute value difference to a score between zero (low similarity, different reinforcement histories in context) and one (strong similarity, similar reinforcement histories in context).

If the agent has learned the value of choosing comparison stimulus a_t in the presence of sample s_t and cue c_b , and the derivation tables specify that c_t mutually entails a relation c^*_t (e.g., in Fig. 2, the table specifies that a sameness relation 'A1 same as B1' mutually entails a sameness relation 'B1 same as A1'), then it can estimate the strength of the mutually entailed relation as:

$$rel_t(c^*_t, a_t, s_t) = rel_t(c_t, s_t, a_t) \quad (5)$$

The agent can then use that relation to compute the value for the associated derived relational response, $Q^d(c^*_b, a_b, s_b)$, as follows:

$$Q^d_t(c^*_t, a_t, s_t) = Q_t(c_t, a_t, s_t) \cdot rel_t(c^*_t, a_t, s_t) \quad (6)$$

This derived value-estimate allows the agent to respond as if the reversed relation (select s_t in the presence of a_t and c_t) had been reinforced, modulated by the strength of the original relation as estimated from prior experience.

The agent can derive combinatorially entailed relations in a similar fashion, now combining estimates for two relations (e.g., the relation between the sample stimulus and another stimulus $s1$, $s_t \sim s1$, and between that same stimulus and a comparison stimulus, $s1 \sim a_t$) as follows:

$$rel_{t+1}(c_t, a_t, s_t) = \frac{1}{2} [rel_t(c_x, s_t, s1) + rel_t(c_y, s1, a_t)] \quad (7)$$

Where $s1$ can be any stimulus that is related to both the sample and comparison stimuli presented on trial t , and c_x and c_y are the contextual cues in the presence of which the stimuli are related. As with mutual entailment, this derived relation can be used to compute an estimate value for the associated derived relational response, $Q^d(c_b, s_b, a_b)$:

$$Q^d_t(c_t, a_t, s_t) = \frac{1}{2} [Q_t(c_x, s_t, s1) + Q_t(c_y, s1, a_t)] \cdot rel_{t+1}(c_t, a_t, s_t) \quad (8)$$

To clarify these derivations, imagine an agent that has completed the Steele and Hayes (1991) training phase (i.e., repeated presentation of Trials 1 to 6 in Fig. 1, allowing for sufficient exploration of the different response options). Starting the test phase, it encounters Trial 7, which presents cue S (i.e., same), sample stimulus B1, and comparison stimuli B2, A1 and B3. B1 has never before been presented as the sample stimulus in this (or any) context, so the agent has no estimated values for selecting any of the comparison stimuli in the presence of this cue-sample pair. Instead, it resorts to its hardcoded knowledge of relational derivation (i.e., the derivation tables, Fig. 2) to check from which source relation(s) the currently cued relation could be derived (e.g., what relation mutually entails the relation cued on the current trial), and checks if it has any knowledge of relations between the sample and comparison stimuli that would fit that derivation (e.g., if I have evidence that A1 is the same as B1, I can derive that B1 is the same as A1, Equation (5)). Provided sufficient exposure to the training phase contingencies, the agent should have learned that in the presence of a cue SAME, stimuli A1 and B1 are strongly related [i.e., $rel(S, A1, B1) \approx 1$], and should expect reinforcement for selecting B1 in the presence of sample A1 and cue S [i.e., positive $Q(S, A1, B1)$]. Using its hardcoded knowledge in combination with the aforementioned relation and stimulus-action values, it can compute (Equation (6)) the value for the derived response [select A1 in the presence of B1 and SAME, $Q^d(S, B1, A1)$]. Similarly, on Trial 17, the agent is presented with cue S, sample B1, and comparison stimuli C1, C2, C3, which it has not experienced before. In addition to having learned about the relation (given cue S) between B1 (the sample) and A, it should have also learned about the relation (given cue S) between A and C1 (one of the comparison stimuli), and it knows that combinatorial entailment of those relations allows for the derivation of the B1-C1 (given cue S) relation (i.e., the derivation table for combinatorial entailment in Fig. 2 indicates that two sameness relations with a common element, here A1, afford the derivation of another sameness relation between the other elements, here B1 and C1). Using Equations (7) and (8), it computes the estimated value of selecting C1 for sample B1 and cue S [i.e., $Q^d(S, B1, C1)$]. On every trial, it tries

derivations for all presented comparison stimuli, but only the correct derived relations should lead to high values (provided sufficient exposure to the training contingencies with deterministic feedback). To make a decision, the derived values are inserted into the softmax function (Equation (3)), along with any other learned estimates for sample-comparison pairs on this trial.

5. Results

5.1. Behavioral data

We first discuss the results of the human participants before providing the model simulation results. Data from all participants were analyzed. Overall, participants learned baseline relations quickly, and on average responded correctly on 30.21 out of 36 trials in the last training block (27 out of 72 participants did not make a single error in this last training block). Participants' cumulative accuracy throughout the training phase is plotted in Fig. 4A, illustrating that the majority of participants catch on to the contingencies relatively quickly and then proceeds nearly flawlessly (straight diagonal lines at the top). However, a minority of the sample took longer to learn (curves straighten out later in training) or failed to do so altogether (curves do not straighten out).

On test trials (i.e., without feedback) assessing those same baseline relations, participants on average responded correctly on 29.47 out of 36 trials ($SD = 8.50$; Fig. 3, left). A one-tailed, one-sample t -test indicated that this performance was significantly higher than chance-level performance: $t(71.00) = 17.36, p < .001$, Cohen's $d = 2.05$, 95% confidence interval (CI) = [1.47, 2.63]. Participants' accuracy reproducing sameness relations was slightly higher (mean accuracy = 10.33 out of 12 trials, $SD = 2.96$; 59 participants reached preregistered 75% accuracy criterion, binomial probability of passing by chance < 0.01) than for difference relations (mean accuracy = 9.64 out of 12, $SD = 3.25$; 51

participants reached 75% accuracy) and opposition relations (mean accuracy = 9.5 out of 12, $SD = 3.33$; 52 participants reached 75% accuracy).

On trials assessing mutual entailment (i.e., reversals of trained relations), participants on average responded correctly on 28.96 out of 36 trials (Fig. 3, middle). A one-sample t -test showed that this was significantly higher than chance-level: $t(71.00) = 16.25, p < .001, d = 1.91$, 95% $CI = [1.35, 2.48]$. Differences in performance for different relations were negligible: participants' mean accuracy reversing sameness relations was 10.47 out of 12 trials ($SD = 2.57$; 61 participants reached 80% accuracy), 10.60 out of 12 for reversing difference relations ($SD = 2.25$; 60 participants reached 75% accuracy) and 10.49 out of 12 for reversing opposition relations ($SD = 2.46$; 58 participants reached 75% accuracy). In the combinatorial entailment block performance dropped to a mean accuracy of 7.03 out of 12 test trials ($SD = 2.96$; Fig. 3, right), and only 21 reached seventy-five percent accuracy. This performance was, however, still significantly higher than chance-level: $t(71.00) = 8.68, p < .001, d = 1.02$, 95% $CI = [0.52, 1.52]$. Given that different relations can be combined together to derive new relations, we did not do relation-based analyses for this block.

5.2. Model simulations

To demonstrate the validity of our computational model, we took the procedure as experienced by one of the participants (one specific randomized trial order) and looped the model through it with different parameter settings. We included four versions of the model: (1) a slow, noisy learner ($\alpha = .15, \tau = 2$), (2) a fast, noisy learner ($\alpha = .75, \tau = 2$), (3) a slow, more deterministic learner ($\alpha = .15, \tau = 0.15$) and (4) a fast, more deterministic learner ($\alpha = .75, \tau = 0.15$). Each version was looped through the task twenty times, after which we took an average to account for variation that is introduced by the softmax choice-function

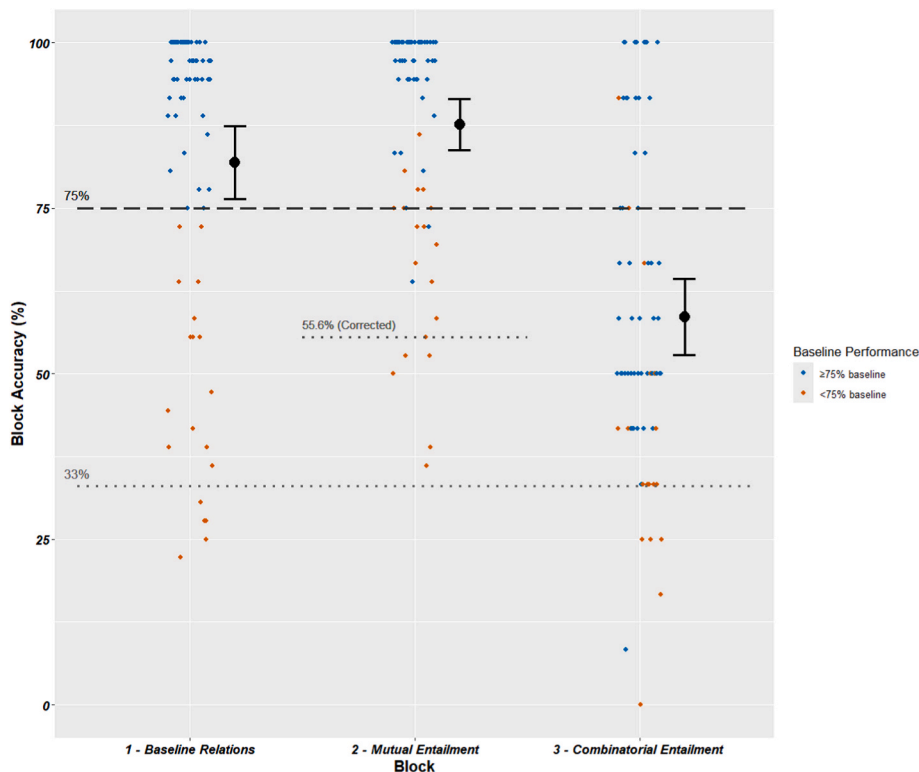


Fig. 3. Participants' performance in the baseline and derived responding test blocks.

Note. The chance-level (33%) and preregistered accuracy criterion (75%) are indicated in grey. Because of the programming error mentioned in the caption of Fig. 1, the chance-level in the mutual entailment block was slightly higher than intended (55.6%), as indicated by the shorter dotted line. Participants are grouped based on performance on baseline relations, illustrating that derived relational responding performance depends on the acquisition of baseline responding.

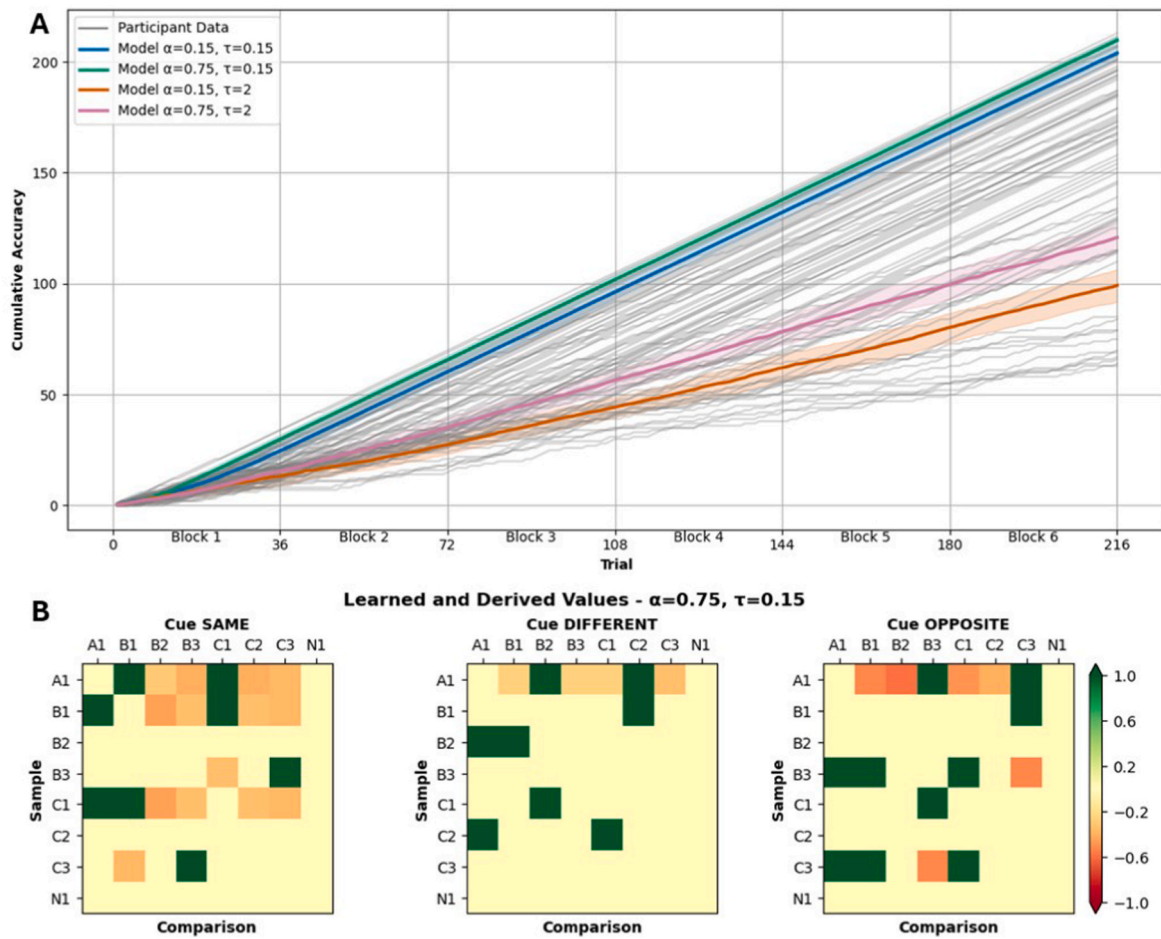


Fig. 4. Participant and model learning trajectories in training and model learned and derived response values.

Note. (A) Participant learning curves are plotted in grey, while curves for different model versions are colored. Shaded areas around model learning curves represent the standard deviation of each model's accuracy throughout training. (B) Each heatmap illustrates the estimated values for selecting a given comparison stimulus (columns) for a particular sample stimulus (rows) and contextual cue (SAME, DIFFERENT or OPPOSITE). Top row represents learned values, other rows represent derived values.

(noise can lead to different options being selected, resulting in different learning trajectories). Fig. 4A shows the learning performance of each agent throughout the training blocks, overlaid over the cumulative accuracy of the sample of human participants throughout training. The plot illustrates how different parameter settings may capture individual variety in human behavior: the models with low decision noise more closely model the better performing participants (with the version with a higher learning rate reaching peak performance sooner), while the models with more noise more closely resemble the participants that have difficulties learning the contingencies (or were responding carelessly or randomly). In the test phase, performance differs for the different model versions (see Supplementary Materials for detailed simulation results). The models with low decision-noise ($\tau = 0.15$) perform best: the model with a higher learning rate performs the test phase flawlessly (with the occasional error due to stochastic decisions), and thus even outperforms most human participants), while the slower-learning model passes the baseline and mutual entailment tests, but performs less consistently in the combinatorial entailment test. Both models with high decision noise ($\tau = 2$) perform inconsistently across all test trials, be it above chance-level.

Fig. 4B illustrates the same model's learned and derived (contextualized) stimulus-response values as a heatmap, illustrating that the agent successfully discriminates patterns of relational responding as a function of the presented cue. It also illustrates the generativity of AARR, with the top row of each of the heatmaps representing the learned values (i.e.,

select a given stimulus for sample A and a particular cue), and the other rows representing derived values. Note that not all possible derived relational responses (see the Derived Network in Fig. 1) have an estimated value assigned to them. This results from the fact that the model only derives values for options presented on a given trial, and not all derived relations are tested. Note that in principle, the agent is capable of deriving all relations possible derived relations (Fig. 1) using the derivation tables (Fig. 2), provided sufficient exposure to the training contingencies (depending on parameter settings).

To get a better illustration of the potential utility of a model like this, we also ran the fast-learning, low-noise ($\alpha = .75, \tau = 0.5$) version of the model through each participant's data. By using the participant's choices, and the feedback they received on them, as input for the model's learning mechanisms, we can test whether they are suitable for predicting human behavior in this task. Specifically, we can on every trial compare the participant's chosen response with the response that the model considers to be the most likely to be reinforced based on prior experience (i.e., the probabilities returned by the softmax function, Equation (3)), and then calculate how well each model predicts each participant's behavior. Across the full sample, on average the model correctly predicts 73.2% of participant's choices (both training and test trials), which is higher than chance-level (33% or 55% in the mutual entailment test). This illustrates the potential of our approach for predicting human behavior. Note that the model can be further adapted to better model particular subgroups of participants.

6. Discussion

Most behavioral experiments in the RFT literature (including Steele & Hayes, 1991) have studied how adult participants bring these previously abstracted, contextually controlled patterns of AARRing to bear in novel contexts (e.g., establishing novel contextual cues through non-arbitrary relational responding training as in the Steele & Hayes, 1991, pretraining) and on novel stimuli (e.g., the network of arbitrary relations trained by Steele & Hayes, 1991), which can be conceived of as short-term learning processes separate from the long-term development of AARR (in line with cognitive accounts of shorter- and longer-term learning processes, e.g., Botvinick et al., 2009; Collins, 2015; and computer science, e.g., Botvinick et al., 2019). Each learning process can be a target for researchers to try and simulate with computational models and might provide unique knowledge. As a first step in this process, we set out to model how adult participants can apply abstract patterns of AARR to novel stimuli in the presence of known contextual cues.

Our approach combines computational RL (Sutton & Barto, 1998) with hardcoded knowledge of relation derivation and contextual cues. In doing so, our artificial agent is able to use reinforcement for responding in a MTS procedure as information to learn about contextually defined stimulus relations. While traditional RL models are perfectly capable of learning how to complete the arbitrary training procedure with reinforcement, they would perform at chance-level in a test phase involving requiring derived responding that was never reinforced before. Our model, however, combines learned stimulus relations (a function of reinforcement in MTS) with hardcoded knowledge of contextual cues and relational derivation to derive which responses to make in the test phase. As such, it serves as an idealized model of adult participants in our conceptual replication of the Steele and Hayes (1991), which again demonstrated the arbitrary, flexible and generative nature of AARR. The model also lends support to RFT's claims that contextual control and abstract relational knowledge are required for AARR.

Because our model demonstrates not only equivalence responding, but also difference and opposition responding (or any other relation or pattern encoded in its hardcoded relational knowledge) in MTS, we consider it an extension of past work on modelling stimulus equivalence (cf. *supra*, see Tovar et al., 2022; for a review), and a contribution to the still scarce literature on computational modelling of AARR. It is, however, limited in multiple respects. First of all, the model depends on hardcoded relational knowledge provided by the researcher and can only respond to those relations and contextual cues it 'knows' (i.e., those that are included in the derivation tables). In its current form, it cannot scale up beyond relatively simple procedures (i.e., typical behavioral experiments with only a few discrete stimuli) or to more complex forms of AARRing (analogy, relational networks). We therefore think of the model as an idealized (i.e., unbound by considerations about cognitive or neural implementation) model of typical adult participants AARRing in behavioral experiments. Further developments, namely constructing computational models of the other learning processes described above, are required to model more complex forms of AARRing in more ecologically valid scenarios.

While limited, we do believe our work has already been fruitful in forcing us to scrutinize RFT and inspiring future modelling research directions. Furthermore, the development of the hardcoded knowledge that the model uses to derive relations, inspired by Allen's (1983) work, also has applications beyond the current model. As Allen argued, such tables can be used for automating certain tasks (in Allen's case, reasoning about temporal sequences of events) with computers. The tables illustrated in Fig. 2 can be expanded to include any number of relations (e.g., comparison, hierarchy, temporal, deictic, and so on), to allow automated reasoning about them. Potentially useful applications for researchers interested in studying relational reasoning are the automated construction of task procedures that have repetitive trial structures (e.g., MTS-procedures or syllogistic reasoning problems), or plotting complex relational networks (e.g., the networks plotted in

Fig. 1). A paper describing these applications in more detail is currently being prepared.

We plan to continue this work by modelling the other short-term learning process present in many behavioral experiments on AARR: the acquisition of contextual cue functions from non-arbitrary relational responding training. Planning for this next step has already identified several areas of the theory that require more specification, or more empirical research. Perhaps surprisingly, the establishment of contextual cues from non-arbitrary relational responding has not been studied systematically, but is often a preliminary step to study how newly established cues influence relational behavior (Delabie et al., 2022; e.g., Steele & Hayes, 1991; Whelan & Barnes-Holmes, 2004; Dymond et al., 2007). While the non-arbitrary cue pretraining phase in such studies can be thought of as an experimental analogue of the long-term abstraction of generalized patterns of AARRing, it is likely that adult participants in such studies have already acquired these abstractions and are rather performing the more trivial task of matching the to-be-established cues to known patterns of relational responding based on experience. Computationally modelling the full extent of the Steele and Hayes procedure would thus require an additional mechanism that allows the agent to perform this matching task. Such a model could then potentially be used to make predictions about open questions in the literature. For instance, little is known about how much MET is required for new cues to be established (i.e., how many different instances of a relational response must be reinforced in the presence of a cue for it to acquire a contextual cue function), or the moderators of that learning process (do contextual control functions extinguish; what are the determinants of the extent of control exerted by a cue).

Another future direction for this work is to extend the model to allow it to deal with multiple stimulus functions concurrently. In our work, we have only taken into account reinforcement (or punishment), because reinforcement (or punishment) for selecting a particular stimulus is the only stimulus function that 'is transformed' in the Steele and Hayes (1991) procedure. Transformation of stimulus function is a crucial aspect of RFT account of AARR, however, and a good computational model of AARR should be capable of representing and responding to multiple stimulus functions. This also highlights a broader limitation of the heavy reliance on MTS procedures in RFT literature. Typically, the MTS procedure merely serves to train up the arbitrary relations along which stimulus functions are to be transformed, while the test for transformation of stimulus function only occurs (once) afterwards (e.g., Dymond et al., 2008; Roche & Barnes, 1997). Without proper procedures to assess the transformation of multiple stimulus functions, a crucial aspect of relational responding, we also cannot properly evaluate our computational models of AARR. Function transformation tasks have recently been developed (see Finn & De Houwer, 2021; Finn et al., 2023), which allow researchers to test transformations of multiple stimulus functions on a trial-by-trial basis. Such procedures would be an interesting test ground for a future version of our model. Finally, the ability to deal with multiple stimulus functions will also be a necessary component for modelling the long-term development of AARR, which involves abstraction from MET involving many different stimulus functions. We therefore expect our efforts modelling the short-term application of AARR will be informative towards our ultimate goal of modelling of its long-term development.

In summary, we have presented a partially hardcoded model of AARR as observed in a conceptual replication of the Steele and Hayes (1991) study. Our behavioral results once again demonstrate the arbitrary, flexible and generative nature of AARR, and the model simulations provide further support for some of RFT's core claims. While the model presented here is obviously limited and idealized, we believe it still represents a contribution to a fledgling literature, in that it highlights aspects of the theory that are conceptually underdeveloped or require more empirical research, and by raising new questions. We are confident that continuing this line of research can provide much more of such knowledge.

CRedit authorship contribution statement

Matthias Raemaekers: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Martin Finn:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Jan De Houwer:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Author note

MR, MF, JDH, Department of Experimental Clinical and Health Psychology, Ghent University. The research presented in this paper was supported by grant BOF22/MET_V/002 - 01M00209 of Ghent University to Jan De Houwer, a fundamental research grant (11M0323N) of the Flemish Research Foundation to Matthias Raemaekers and a Flemish Research Foundation Grant (12A2B26N) to Martin Finn. Data and materials are available at the Open Science Framework (https://osf.io/8xbnj/?view_only=f443851100f24b07b3b6e2a95ac6f7f2).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Matthias Raemaekers reports financial support was provided by Research Foundation Flanders. Jan De Houwer reports financial support was provided by Bijzonder Onderzoeksfonds. Martin Finn reports financial support was provided by Research Foundation Flanders. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jcbs.2026.101002>.

References

- Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11), 832–843. <https://doi.org/10.1145/182.358434>
- Barnes, D., & Hampson, P. J. (1993). Stimulus equivalence and connectionism: Implications for behavior analysis and cognitive science. *Psychological Record*, 43(4), 617–638. <https://doi.org/10.1007/bf03395903>
- Barnes, D., Hegarty, N., & Smeets, P. M. (1997). Relating equivalence relations to equivalence relations: A relational framing model of complex human functioning. *The Analysis of Verbal Behavior*, 14(1), 57–83. <https://doi.org/10.1007/BF03392916>
- Barnes-Holmes, D., Barnes-Holmes, Y., & McEnteggart, C. (2020). Updating RFT (more field than frame) and its implications for process-based therapy. *Psychological Record*, 70(4), 605–624. <https://doi.org/10.1007/s40732-019-00372-3>
- Barnes-Holmes, Y., Barnes-Holmes, D., Smeets, P. M., Strand, P., & Friman, P. (2004). Establishing relational responding in accordance with more-than and less-than as generalized operant behavior in young children. *International Journal of Psychology and Psychological Therapy*, 4, 531–558.
- Belisle, J., & Dixon, M. R. (2020). Relational density theory: Nonlinearity of equivalence relating examined through higher-order volumetric-mass-density. *Perspectives on Behavior Science*, 43(2), 259–283. <https://doi.org/10.1007/s40614-020-00248-w>
- Binz, M., & Schulz, E. (2023). Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.
- Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3), 262–280. <https://doi.org/10.1016/j.cognition.2008.08.011>
- Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5), 408–422. <https://doi.org/10.1016/j.tics.2019.02.006>
- Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, 107(4), 603–616. <https://doi.org/10.1016/j.neuron.2020.06.014>
- Carrillo, A., & Betancort, M. (2023). Differences of training structures on stimulus class formation in computational agents. *Multimodal Technologies and Interaction*, 7(4), 39. <https://doi.org/10.3390/mti7040039>
- Carrillo, A., & Betancort, M. (2024). Testing stimulus equivalence in transformer-based agents. *Future Internet*, 16(8), 289. <https://doi.org/10.3390/fi16080289>
- Cassidy, S., Roche, B., & Hayes, S. C. (2011). A relational frame training intervention to raise intelligence quotients: A pilot study. *Psychological Record*, 61(2), 173–198. <https://doi.org/10.1007/bf03395755>
- Cullinan, V. A., Barnes, D., Hampson, P. J., & Lyddy, F. (1994). A transfer of explicitly and nonexplicitly trained sequence responses through equivalence relations: An experimental demonstration and connectionist model. *Psychological Record*, 44(4), 559–585. <https://doi.org/10.1007/bf03395144>
- Delabie, M., Cummins, J., Finn, M., & De Houwer, J. (2022). Differential crel and Cfunc acquisition through stimulus pairing. *Journal of Contextual Behavioral Science*, 24, 112–119. <https://doi.org/10.1016/j.jcbs.2022.03.012>
- Dixon, M. R. (2014). *PEAK relational training system: Evidence-based autism assessment and treatment*. Shawnee Scientific Press.
- Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the experimental analysis of behavior*, 62(3), 331–351. <https://doi.org/10.1901/jeab.1994.62-331>
- Dougher, M. J., & Markham, M. R. (1994). Stimulus equivalence, functional equivalence and the transfer of function. *Behavior analysis of language and cognition*, 71–90.
- Duran, N., Brown, F. J., & Cowell, D. (2026). *Modelling human relational reasoning in graph neural networks*. Available at: SSRN 6156081.
- Dymond, S., & Barnes, D. (1994). A transfer of self-discrimination response functions through equivalence relations. *Journal of the Experimental Analysis of Behavior*, 62(2), 251–267. <https://doi.org/10.1901/jeab.1994.62-251>
- Dymond, S., & Barnes, D. (1995). A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness, more than, and less than. *Journal of the Experimental Analysis of Behavior*, 64(2), 163–184. <https://doi.org/10.1901/jeab.1995.64-163>
- Dymond, S., & Barnes, D. (1996). A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness and opposition. *Psychological Record*, 46(2), 271.
- Dymond, S., Roche, B., Forsyth, J. P., Whelan, R., & Rhoden, J. (2008). Derived avoidance learning: Transformation of avoidance response functions in accordance with same and opposite relational frames. *The Psychological Record*, 58(2), 269–286. <https://doi.org/10.1007/bf03395615>
- Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, 35(4), 639–665. [https://doi.org/10.1016/s0005-7894\(04\)80013-3](https://doi.org/10.1016/s0005-7894(04)80013-3)
- Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A Post-Skinnerian account of human language and cognition*. Springer Science & Business Media. <https://doi.org/10.1007/b108413>
- Hayes, S. C., Law, S., Assemi, K., Falletta-Cowden, N., Shamblin, M., Burleigh, K., ... Smith, P. (2021). Relating is an operant: A fly over of 35 years of RFT research. *Perspectivas em Análise do Comportamento*, 12(1), 5–32. <https://doi.org/10.18761/pac.2021.v12.rft.02>
- Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2011). *Acceptance and commitment therapy: The process and practice of mindful change*. Guilford press.
- Johansson, R., Hammer, P., & Lofthouse, T. (2024). *Functional equivalence with nars*. *arXiv preprint arXiv:2405.03340*.
- Johansson, R., Hammer, P., & Lofthouse, T. (2025). Arbitrarily applicable same/opposite relational responding with NARS. In *International conference on artificial general intelligence* (pp. 314–324). Cham: Springer Nature Switzerland.
- Johansson, R., & Lofthouse, T. (2023). Stimulus equivalence in nars. In *International conference on artificial general intelligence* (pp. 158–166). Cham: Springer Nature Switzerland.
- Johansson, R., Lofthouse, T., & Hammer, P. (2022). Generalized identity matching in nars. In *International conference on artificial general intelligence* (pp. 243–249). Cham: Springer International Publishing.
- Lipkens, R., & Hayes, S. C. (2009). Producing and recognizing analogical relations. *Journal of the experimental analysis of behavior*, 91(1), 105–126. <https://doi.org/10.1901/jeab.2009.91-105>
- Lipkens, R., Hayes, S. C., & Hayes, L. J. (1993). Longitudinal study of the development of derived relations in an infant. *Journal of Experimental Child Psychology*, 56(2), 201–239. <https://doi.org/10.1006/jecp.1993.1032>
- Luciano, C., Becerra, I. G., & Valverde, M. R. (2007). The role of multiple-exemplar training and naming in establishing derived equivalence in an infant. *Journal of the experimental analysis of behavior*, 87(3), 349–365. <https://doi.org/10.1901/jeab.2007.08-06>
- Lyddy, F., & Barnes-Holmes, D. (2007). Stimulus equivalence as a function of training protocol in a connectionist network. *Journal of Speech - Language Pathology and Applied Behavior Analysis*, 2(1), 14. <https://doi.org/10.1037/h0100204>
- Lyddy, F., Barnes-Holmes, D., & Hampson, P. J. (2001). A transfer of sequence function via equivalence in a connectionist network. *Psychological Record*, 51(3), 409–428. <https://doi.org/10.1007/bf03395406>
- McHugh, L., Barnes-Holmes, Y., & Barnes-Holmes, D. (2004). Perspective-taking as relational responding: A developmental profile. *Psychological Record*, 54(1), 115–144. <https://doi.org/10.1007/bf03395465>
- McIlvane, W. J. (2003). A stimulus in need of a response: A review of relational frame theory: A post-skinnerian account of Human language and cognition. *The Analysis of Verbal Behavior*, 19(1), 29–37. <https://doi.org/10.1007/bf03392980>
- Mofrad, A. A., Yazidi, A., Hammer, H. L., & Arntzen, E. (2020). Equivalence projective simulation as a framework for modeling formation of stimulus equivalence classes. *Neural Computation*, 32(5), 912–968. https://doi.org/10.1162/neco_a_01274

- Ninness, C., & Ninness, S. K. (2020). Emergent virtual analytics: Modeling contextual control of derived stimulus relations. *Behavior and Social Issues*, 29(1), 119–137. <https://doi.org/10.1007/s42822-020-00032-0>
- Ninness, C., Ninness, S. K., Rumph, M., & Lawson, D. (2018). The emergence of stimulus relations: Human and computer learning. *Perspectives on Behavior Science*, 41(1), 121–154. <https://doi.org/10.1007/s40614-017-0125-6>
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3), 139–154. <https://doi.org/10.1016/j.jmp.2008.12.005>
- O'Hora, D., Barnes-Holmes, D., & Stewart, I. (2014). Antecedent and consequential control of derived instruction-following. *Journal of the experimental analysis of behavior*, 102(1), 66–85. <https://doi.org/10.1002/jeab.95>
- O'Sullivan, J., Jackson Brown, F., & Ray, O. (2025). Elucidating simulated equivalence responding through dynamic visualization of structural connectivity and relational density. *Frontiers in Artificial Intelligence*, 8, Article 1618678. <https://doi.org/10.3389/frai.2025.1618678>
- O'Connor, M., Farrell, L., Munnely, A., & McHugh, L. (2017). Citation analysis of relational frame theory: 2009–2016. *Journal of Contextual Behavioral Science*, 6(2), 152–158. <https://doi.org/10.1016/j.jcbs.2017.04.009>
- O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, 1, 94–100. <https://doi.org/10.1016/j.cobeha.2014.10.004>
- O'Hora, D., Barnes-Holmes, D., Roche, B., & Smeets, P. (2004). Derived relational networks and control by novel instructions: A possible model of generative verbal responding. *Psychological Record*, 54(3), 437–460. <https://doi.org/10.1007/bf03395484>
- OpenAI. (2024). *Learning to reason with LLMs*. <https://openai.com/index/learning-to-reason-with-llms/>.
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and brain sciences*, 31(2), 109–130. <https://doi.org/10.1017/s0140525x08003543>
- Perez, W. F., Harte, C., Barnes-Holmes, D., Gomes, C. T., Mohor, B., & de Rose, J. C. (2023). Generalized contextual control based on nonarbitrary and arbitrary transfer of stimulus functions. *Journal of the Experimental Analysis of Behavior*, 119(3), 448–460. <https://doi.org/10.1002/jeab.839>
- Roche, B., & Barnes, D. (1997). A transformation of responsively conditioned stimulus function in accordance with arbitrarily applicable relations. *Journal of the Experimental Analysis of behavior*, 67(3), 275–301. <https://doi.org/10.1901/jeab.1997.67-275>
- Roche, B., Barnes-Holmes, D., Barnes-Holmes, Y., Smeets, P. M., & McGeady, S. (2000). Contextual control over the derived transformation of discriminative and sexual arousal functions. *Psychological Record*, 50(2), 267–291. <https://doi.org/10.1007/bf03395356>
- Steele, D., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior*, 56(3), 519–555. <https://doi.org/10.1901/jeab.1991.56-519>
- Stewart, I., Barnes-Holmes, D., Hayes, S. C., & Lipkens, R. (2001). Relations among relations: Analogies, metaphors, and stories. In *Relational frame theory: A Post-Skinnerian account of human language and cognition* (pp. 73–86). Boston, MA: Springer US. https://doi.org/10.1007/0-306-47638-x_4
- Stewart, I., Barnes-Holmes, D., & Roche, B. (2004). A functional-analytic model of analogy using the relational evaluation procedure. *Psychological Record*, 54(4), 531–552. <https://doi.org/10.1007/bf03395491>
- Stewart, I., Barnes-Holmes, D., Roche, B., & Smeets, P. M. (2002). A functional-analytic model of analogy: A relational frame analysis. *Journal of the Experimental Analysis of Behavior*, 78(3), 375–396. <https://doi.org/10.1901/jeab.2002.78-375>
- Subramanian, A., Chitlangia, S., & Baths, V. (2022). Reinforcement learning and its connections with neuroscience and psychology. *Neural Networks*, 145, 271–287. <https://doi.org/10.1016/j.neunet.2021.10.003>
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. The MIT Press.
- Tovar, A. E., & Chávez, A. T. (2012). A connectionist model of stimulus class formation with a yes/no procedure and compound stimuli. *Psychological Record*, 62(4), 747–762. <https://doi.org/10.1007/bf03395833>
- Tovar, A. E., Torres-Chávez, Á., Mofrad, A. A., & Arntzen, E. (2023). Computational models of stimulus equivalence: An intersection for the study of symbolic behavior. *Journal of the Experimental Analysis of Behavior*, 119(2), 407–425. <https://doi.org/10.1002/jeab.829>
- Vernucio, R. R., & Debert, P. (2016). Computational simulation of equivalence class formation using the go/no-go procedure with compound stimuli. *Psychological Record*, 66(3), 439–449. <https://doi.org/10.1007/s40732-016-0184-1>
- Wang, P. (2006). *Rigid flexibility: The logic of intelligence*. Netherlands: Dordrecht: Springer.
- Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3), 279–292. <https://doi.org/10.1007/bf00992698>
- Whelan, R., & Barnes-Holmes, D. (2004). The transformation of consequential functions in accordance with the relational frames of same and opposite. *Journal of the Experimental analysis of Behavior*, 82(2), 177–195. <https://doi.org/10.1901/jeab.2004.82-177>
- Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, 8, Article e49547. <https://doi.org/10.7554/elife.49547>