Scientific progress in AGI from the perspective of contemporary behavioral psychology

Robert Johansson[1,2]

[1] Department of Computer and Information Science, Linköping University, Sweden
[2] Department of Psychology, Stockholm University, Sweden
robert.johansson@liu.se

Abstract. While the field of Artificial General Intelligence (AGI) continues to develop, it seems to be a need for continued progress in terms of developing AGI as a science. In this paper, we discuss scientific progress in AGI from the perspective of behavioral psychology. We provide examples of psychological experiments that seem potentially useful for AGI research. First, we show experiments that demonstrate various cognitive capabilities. Then, in line with contemporary behavioral psychology research, we examine how the terms coherence, complexity and levels of derivation can be used to study the dynamics of complex responding. Finally, we discuss experiments that are uniquely possible for AGI. Future implications for the AGI field are discussed.

Keywords: Science · Behavioral psychology · Relational Frame Theory.

1 Introduction

There has been a call for theoretical development in the field of Artificial General Intelligence (AGI) [14]. As suggested by Wang [16], theoretical development in the AGI field, is very much dependent on the definition of "artificial intelligence" in itself, as a chosen definition will influence the path of a particular research project. As AGI continues to develop, there is a need for multiple perspectives and continued theoretical development within these.

We have previously argued for the value of contemporary behavioral psychology in the field of AGI [7]. At the heart of our argument is the fact that a behavioral psychology definition of learning seems compatible to Wang's definition of intelligence as *adaptation with insufficient knowledge and resources* [16]. Contemporary behavioral psychology defines learning as *ontogenetic adaptation*, that is, the *adaptation of an individual organism to its environment during the lifetime of the individual* [3].

Given this compatibility between perspectives from AI and psychology, we believe it is fruitful to discuss scientific progress in AGI from the perspective of behavioral psychology. Below, we present a set of experimental setups to illustrate scientific progress in behavioral psychology that could be of value to AGI. Empirical data will also be used to point to potentially relevant AGI research.

2 Match-to-sample experiments

The Match-to-sample (MTS) task is a common experimental setup when studying cognitive behavior from a behavioral psychology perspective. MTS involves the presentation of a single stimulus, often referred to as the sample stimulus. In addition, two or more stimuli are presented, often referred to as the comparison stimuli. The task of the experimental participant is to respond to one of the comparison stimuli (for example by pressing left or right on a keyboard). Sample and comparison stimuli are typically visual (for example pictures, objects or words), but can in principle be of any sensory modality. After a selection has been made by the participant, feedback is provided (typically an indication of *Correct* or *Incorrect*). As a correct response is conditional upon a particular sample in a trial, the term *conditional discrimination learning* is used to describe the type of learning involved in the MTS. This is illustrated in Figure 1.

2.1 Learning conditional discriminations

If $A1$ is presented as sample, with $B1$ and $B2$ as possible choices, feedback can be given so that $B1$ is consistently picked when $A1$ is presented ($A1$, $B1$, and $B2$ being arbitrary symbols). In the text below such experimental setup is denoted as $A1|B1, B2$, and $A1 \rightarrow B1$ is used to denote the learned relation. With a similar setup, $A2 \rightarrow B2$ can be taught with feedback.

In a formal experimental situation, multiples of the four trials $A1|B1, B2$, $A1|B2, B1$, $A2|B1, B2$, and $A2|B2, B1$ are presented as a block of for example 16 trials. A success criteria is set to typically 15 out of 16 trials. If the participant does not pass the success criteria, the block is repeated. The conditional discriminations experiment ends when the participant pass the criteria. Then, the participant can be said to have learned a set of "if-then-relations", like *"If A1 then B1"* and *"If A2 then B2"*.

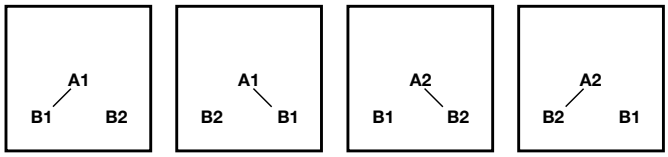

Fig. 1. Learning conditional discriminations in the Match-to-sample task, over the trials $A1|B1, B2$, $A1|B2, B1$, $A2|B1, B2$, and $A2|B2, B1$. If $A1$ and $A2$ is presented, with $B1$ and $B2$ as possible choices, feedback can be given so that $B1$ is consistently picked when $A1$ is presented, and $B2$ when $A2$ is presented ($A1$, $B1$, and $B2$ being arbitrary symbols). $A1 \rightarrow B1$ and $A2 \rightarrow B2$ are used to denote the learned relations.

2.2 Generalized identity matching

Identity matching is a form of concept learning, possible to study using the the match-to-sample experiment. In an identity matching task, a response to a comparison stimuli that is identical to the sample is reinforced. That is, the participant responds to situations such as $A1|A1, A2$, $A1|A2, A1$, $A2|A1, A2$, and $A2|A2, A1$. Hence, the relations $A1 \rightarrow A1$ and $A2 \rightarrow A2$ are learned.

Importantly though, to determine if the matching performance generalizes (transfers to novel situations), the training above is followed by tests with novel stimuli. For example, $A3|A3, A4$, $A3|A4, A3$, $A4|A3, A4$, and $A4|A4, A3$. Typically no feedback is given in this part of the experiment. If a participant pass criteria in the testing phase, this is taken as evidence that the identity concept has been learned. This means that it is the identity relation between sample and comparison that is assumed to control the response, rather than simple conditional discriminations. Hence, an abstract concept of identity has been demonstrated.

Among non-humans, generalized identity matching has been demonstrated with pigeons, monkeys, dolphins and sea lions using visual stimuli, and with rats using olfactory stimuli [12].

2.3 Symmetry

Symmetry is the finding that after a verbally capable experimental participant learns to match samples to comparisons in a conditional discrimination task, the participant will then match the same stimuli when the sample and comparison roles are reversed. For example, if $A1 \rightarrow B1$ and $A2 \rightarrow B2$ are taught as above, a test can be conducted using trials such as $B1|A1, A2$, $B1|A2, A1$, $B2|A1, A2$, and $B2|A2, A1$. To pass the test a participant needs to derive *"If B1 then A1"* given *"If A1 then B1"*.

This is indeed an important cognitive capability, as it can be said to be the smallest example of "derived knowing". To have knowledge derived, without a direct learning experience is undoubtedly a key to advanced cognitive capabilities. In humans, symmetry seems to develop before 24 months of age [9]. There is some evidence that symmetry in infants develop with the help of *multiple-exemplar training* [10]. Among non-humans, the evidence is mixed for symmetry. There are inconsistent results over repeated studies of pigeons, monkeys, and rats [8].

2.4 Stimulus equivalence

Assume that a participant has been trained in four relation using the MTS procedure, $A1 \rightarrow B1$, $B1 \rightarrow C1$, $A2 \rightarrow B2$, and $B2 \rightarrow C2$. Without further training, a verbally able participant will demonstrate an increased probability of not only symmetrical responses, but also transitive responses ($A1 \rightarrow C1$ and $A2 \rightarrow C2$) and *stimulus equivalence* ($C1 \rightarrow A1$ and $C2 \rightarrow A2$). Stimulus equivalence is a behavioral phenomenon that seems to be limited to humans with verbal abilities [17].

2.5 Contextually controlled derived relational responding

A more general version of symmetry is when the derived reversed response is *contextually controlled*. For example, if a verbal human learns that *"A1 is more than B1"*, then the person derives that *"B1 is less than A1"*. This can be studied in a match-to-sample experiment, where an additional symbol has been pre-trained to denote the *MORE/LESS* relation. This symbol functions as a *contextual cue* for which relation to learn. For example, in the presence of the *MORE* cue, learning to choose $B1$ when $A1$ is the sample, the relation *"A1 is more than B1"*, will be taught. Similarly, when the *LESS* cue is present, learning to choose $B2$ in relation to $A1$, implies learning of *"A1 is less than B2"*.

Given that these relations have been learned, the reversed relations can be tested for. This was conducted in the study by O'Hora and co-authors [11]. Figure 2 illustrates the training and testing procedures from that study. The authors did also train and test derived reversed performance on *SAME* and *OPPOSITE* relations. As the participants' responses in the study were different depending on which relational cue was used, the arbitrary nature of the applications of the relational concepts was evident. Importantly, the participants in the study were adults, that already had a history of using the abstract concepts of *SAME*, *OPPOSITE* and *MORE/LESS*.

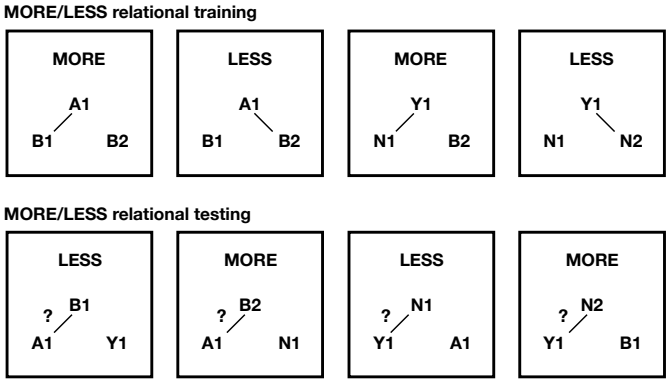

Fig. 2. The relations trained (top) and tested (bottom) by O'Hora and co-authors [11]. In the actual experiment, the symbols $A1$, $B1$, etc, were nonsense symbols. The words *MORE/LESS* were not used for the contextual cues. Instead, symbols were pre-trained to be associated with the respective relations. The details of the experiment is outside the scope of this paper, but the test tasks were designed to eliminate alternative explanations for the results.

3 Arbitrarily applicable relational responding

The above responses are instances of *arbitrarily applicable relational responding (AARR)*. We have previously described AARR as a behavioral psychology approach to general intelligence [7]. The contemporary behavioral psychology theory Relational Frame Theory suggest that AARR is a necessity for intelligence and higher-order cognitive tasks [5].

AARR is defined as *abstract response patterns*, that have the properties of *mutual entailment*, *combinatorial entailment* and *transformation of stimulus functions*, that are controlled by *contextual cues* and learned through a history of *multiple exemplar training*. Specific instances of AARR (for example sameness and comparison), are referred to as different types of *relational frames*. In this context, *abstract response patterns* means that it is an application of an abstract concept (as discussed above for generalized identity matching). Regarding *mutual entailment* this can be said the be a generalized form of symmetry, where the bidirectional response might trigger different relations (as in the *MORE/LESS* case). *Combinatorial entailment* is similarly a generalized form of transitivity. *Transformation of stimulus functions* refers to the fact that multiple stimulus functions (e.g. appetative/aversive, perceptual, discriminative) can be transferred or transformed over relational networks. The above are explained in more detail in [7].

The fact that a contextual cue (such as *MORE* or *SAME*) will trigger different responses highlights the *contextual control* aspect of AARR. As mentioned above, the AARR patterns are learned through *multiple exemplar training*. That is, they are trained repeatedly across different situations, where the details vary, and the only thing being invariant is the abstract concept [10].

Children seem to develop *SAME* relations before two years of age [9], and relations beyond equivalence (such as *MORE/LESS* and *OPPOSITE*) around the age of four years [2].

4 Theoretical development and the dynamics of AARR

In our opinion, all of the experiments described above are interesting in an AGI context. The experiments illustrate different forms of cognitive behavior, increasing in difficulty. We believe it is a potential path of scientific progress in AGI to explore these experiments, and many more as suggested by research based on Relational Frame Theory. As AGI aim to build "thinking machines", demonstrating various forms of AARR in AGI systems is indeed important. Still, the research described above is on the form *"Given experience E, will behavior B happen? (yes/no)"*. For example, what experience would an AGI system need to display symmetry in an experimental situation? Or, what kind of multiple exemplar training (if any) would be needed for a system to be able to derive *LESS* relations given training in *MORE* relations? There is simply a limit to the applicability of such potential findings, that merely demonstrates the existence of an advanced behavior in a certain context.

Recent conceptual development in Relational Frame Theory has suggested moving from demonstrating various forms of AARR to studying the *dynamics of AARR* [1]. The authors suggest focusing on the following features of AARR: *coherence*, *complexity*, and *levels of derivation*. That is, what experiences give rise to a variation in these features of AARR?

Coherence Someone can be said to respond *coherently* when the response and all derived relations are consistent with what the individual had learned previously [17]. More formally, high coherence indicates that a given pattern of AARR is in line with previous patterns of AARR [1]. For example, the statement *"B is larger than A"* is coherent with the statement *"A is smaller than B"*.

Complexity Relational complexity refers to the "intricacy" of a pattern of AARR [1]. For example in the context of stimulus equivalence (see above), a response $D1 \rightarrow A1$ given a history of four symbols, is less complex than the response $C1 \rightarrow A1$ from the three-node network taught in the example above. In addition, a response *"B1 is less than A1"* given *"A1 is more than B1"*, is more complex than *"D2 is the opposite of C1"* given *"C1 is the opposite of D2"*. This is due to the fact that the *MORE/LESS* response entails two relations, while the *OPPOSITE* only is about one.

Levels of derivation Levels of derivation refers to how well established an AARR response has become [1]. A response is said to be *high in derivation* if it is derived for the very first time (i.e., the response is highly novel). The more times an AARR response is emitted, the lower in derivation the response is.

5 Studying the dynamics of AARR in AGI

One RFT study examined coherence in the context of ambiguous scenarios [13]. The authors found that the participants in the experiment tended to derive coherent relations when presented with an ambiguous relational tasks. We belive this could inspire AGI research, where systems need to act despite being confronted with ambiguity.

Research from Relational Frame Theory on relational complexity seems to indicate that more complex responses takes longer time to emit, than less complex counterparts [6]. In the study by O'Hora mentioned above [11], response times were compared between *SAME/OPPOSITE* responses and *MORE/LESS* responses. The latter took longer time to emit, as predicted by the authors. In the same study, the authors allowed the participants to "practice" the complex responses (i.e., to decrease levels of derivation), which lowered the response time [11]. Both these results from the O'Hora study seem interesting to study in AGI system. For example, we believe that there could be interesting parallels between the AARR features coherence, complexity and levels of derivation, and the uncertainty measures *frequency* and *confidence*, used in Pei Wang's Non-axiomatic Reasoning System (NARS).

6 Scientific progress in AGI design

In the discussion above, we have provided examples of two broad classes of experiments that could be relevant to AGI researchers. First, we provided a set of experiments that could demonstrate increasingly advanced cognitive capabilities, from the perspective of Relational Frame Theory. Scientific progress in AGI could be to uncover what experiences a system would need to have to be able to perform in these experiments. Secondly, we discussed the dynamics of AARR. That is, what experiences would a system need to demonstrate changes in the coherence, complexity, and/or levels of derivation of an AARR response?

All of the above has been about scientific progress given changes in experience (the "nurture" part). We now turn to scientific progress in the design of a system ("nature"). This is research that simply is not possible in human psychology, but indeed is possible in AGI. The example below will be from NARS [15].

NARS is implemented in nine layers, with more advanced capabilities on the upper layers. At each layer, multiple derivation rules exist. One could imagine a scenario where a derivation rule or an entire layer were removed from a NARS system, and that system would be compared to a full NARS system in a standardized experiment (like those described above). If only the full NARS system could succeed in the experiment, then that would be evidence of the need for a certain part of NARS in an experiment. This experimental approach could be varied in very many ways, with different NARS components being removed, and tested in different experimental setups. Such approach could indeed be a foundation for a kind of scientific progress in AGI.

Recently, Patrick Hammer released his ANSNA model, derived from NARS [4]. Hammer demonstrated that ANSNA could learn to do generalized identity matching. Importantly though, he tested a version of ANSNA without capabilities from NARS layer 6 (introducing variables for generalization), which resulted in an ANSNA version that could not perform identity matching. Hence, this is an example of scientific progress in AGI: The generalization functionality implemented with variables, seemed to be necessary for generalized identity matching to happen.

7 Conclusion

In this paper, we have restated the need for scientific progress in the AGI field. We have illustrated the bottom-up theoretical development enabled by conducting psychological experiments. Behavioral psychology in general, and Relational Frame Theory in particular provide an interesting roadmap for AGI researchers, by suggesting a range of interesting experiments. Importantly, AGI by design opens up for experimental manipulation "inside" AGI systems, and not only in their experience. We believe this is a fruitful path for theoretical knowledge to be generated within the AGI field.

Acknowledgements The author would like to thank Arne Jönsson and Sam Thellman at Linköping University for valuable discussions regarding this paper.

References

1. Barnes-Holmes, D., Barnes-Holmes, Y., Luciano, C., McEnteggart, C.: From the irap and rec model to a multi-dimensional multi-level framework for analyzing the dynamics of arbitrarily applicable relational responding. Journal of contextual behavioral science **6**(4), 434–445 (2017)
2. Barnes-Holmes, Y., Barnes-Holmes, D., Smeets, P.M., Strand, P., Friman, P.: 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 (2004)
3. De Houwer, J., Barnes-Holmes, D., Moors, A.: What is learning? on the nature and merits of a functional definition of learning. Psychonomic bulletin & review **20**(4), 631–642 (2013)
4. Hammer, P.: Adaptive neuro-symbolic network agent. In: International Conference on Artificial General Intelligence. pp. 80–90. Springer (2019)
5. Hayes, S.C., Barnes-Holmes, D., Roche, B.: Relational frame theory : a post-Skinnerian account of human language and cognition. Kluwer Academic/Plenum Publishers, New York (2001)
6. Hughes, S., Barnes-Holmes, D., Vahey, N.: Holding on to our functional roots when exploring new intellectual islands: A voyage through implicit cognition research. Journal of Contextual Behavioral Science **1**(1-2), 17–38 (2012)
7. Johansson, R.: Arbitrarily applicable relational responding. In: International Conference on Artificial General Intelligence. pp. 101–110. Springer (2019)
8. Lionello-DeNolf, K.M.: The search for symmetry: 25 years in review. Learning & behavior **37**(2), 188–203 (2009)
9. Lipkens, R., Hayes, S.C., Hayes, L.J.: Longitudinal study of the development of derived relations in an infant. Journal of Experimental Child Psychology **56**(2), 201–239 (1993)
10. Luciano, C., Becerra, I.G., Valverde, M.R.: 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 (2007)
11. O'Hora, D., Roche, B., Barnes-Holmes, D., Smeets, P.M.: Response latencies to multiple derived stimulus relations: Testing two predictions of relational frame theory. The Psychological Record **52**(1), 51–75 (2002)
12. Peña, T., Pitts, R.C., Galizio, M.: Identity matching-to-sample with olfactory stimuli in rats. Journal of the experimental analysis of behavior **85**(2), 203–221 (2006)
13. Quinones, J.L., Hayes, S.C.: Relational coherence in ambiguous and unambiguous relational networks. Journal of the experimental analysis of behavior **101**(1), 76–93 (2014)
14. Wang, P.: Theories of artificial intelligence—meta-theoretical considerations. In: Theoretical Foundations of Artificial General Intelligence, pp. 305–323. Springer (2012)
15. Wang, P.: Non-axiomatic logic: A model of intelligent reasoning. World Scientific (2013)
16. Wang, P.: On defining artificial intelligence. Journal of Artificial General Intelligence **10**(2), 1–37 (2019)
17. Zettle, R.D., Hayes, S.C., Barnes-Holmes, D., Biglan, A.: The Wiley handbook of contextual behavioral science. Wiley Online Library (2016)