From Subsymbolic to Symbolic: A Blueprint for Investigation

Joseph Pober[1,*,†], Michael Luck[1,†] and Odinaldo Rodrigues[1,†]

[1]Department of Informatics, Bush House, King's College London, London WC2B 4BG, UK

Abstract

In this paper, we sketch a framework for integration between subsymbolic and symbolic representations, consisting of a series of layers and mappings between elements across the layers. Each layer corresponds to a particular level of abstraction about phenomena in the environment being observed in the layers below. Through an iterative process, the differences between the elements in successive iterations within a given layer are captured as transformations between the elements and used for identification and recognition of objects as well as prediction and verification of the environment in future iterations. A bridge between the subsymbolic and symbolic levels can be built by successively adding layers at ever more sophisticated levels of abstraction. This approach aims to benefit from subsymbolic learning, while harnessing the abstraction and reasoning powers of classical symbolic AI techniques.

Keywords

neuro-symbolic integration, predicate learning, learning structured representations

1. Introduction

While extremely valuable, especially in domains with an abundance of existing data, such as in game playing and natural language processing, subsymbolic techniques often have serious shortcomings in aspects that can be critical for the development of Artificial General Intelligence, e.g., abstraction, transfer learning, interpretability [1]. In humans, one can argue that these cognitive functions have evolved in tandem with the development of natural language, allowing a deeper understanding of the world around us through the construction of ever more sophisticated abstract *symbolic* models. There are obvious advantages of these symbolic models: they can be communicated between individuals, translated between (formal) languages, be refined and revised, used in different domains, composed, etc.

Classical AI often uses formal languages such as logic to represent the world and to model concepts such as actions and change. While such *reasoning* models overcome the shortcomings of purely subsymbolic approaches, they cannot easily learn from new data, or construct or revise themselves. Combining the power of subsymbolic approaches with the elegance and flexibility of the symbolic ones has huge potential and yet it has proven elusive [2].

The central message of this paper is that the bridge between subsymbolic and symbolic representations needs to be built as a series of layers of *abstractions* mapping artefacts and concepts between the layers, using basic building blocks through which symbols can be constructed from entirely subsymbolic data. The ultimate aim is to build a top-level symbolic layer which can be used to describe phenomena of interest perceived by the lower subsymbolic ones. Our focus is on how to create these building blocks in an unsupervised fashion and to devise mechanisms by which they can be used to generate symbols representing a variety of concepts, e.g., objects, transformations, relations. We illustrate our ideas by considering the well-known game of Pong, which consists of a ball and two paddles. In our setting, this game is perceived through a sequence of still images associated with consecutive snapshots of the game.

NESY 2022: 16th International Workshop on Neural-Symbolic Learning and Reasoning, Cumberland Lodge, Windsor, UK

*Corresponding author.

†These authors contributed equally.

joseph.pober@kcl.ac.uk (J. Pober); michael.luck@kcl.ac.uk (M. Luck); odinaldo.rodrigues@kcl.ac.uk (O. Rodrigues)

00000-0002-0926-2061 (M. Luck); 0000-0001-7823-1034 (O. Rodrigues)

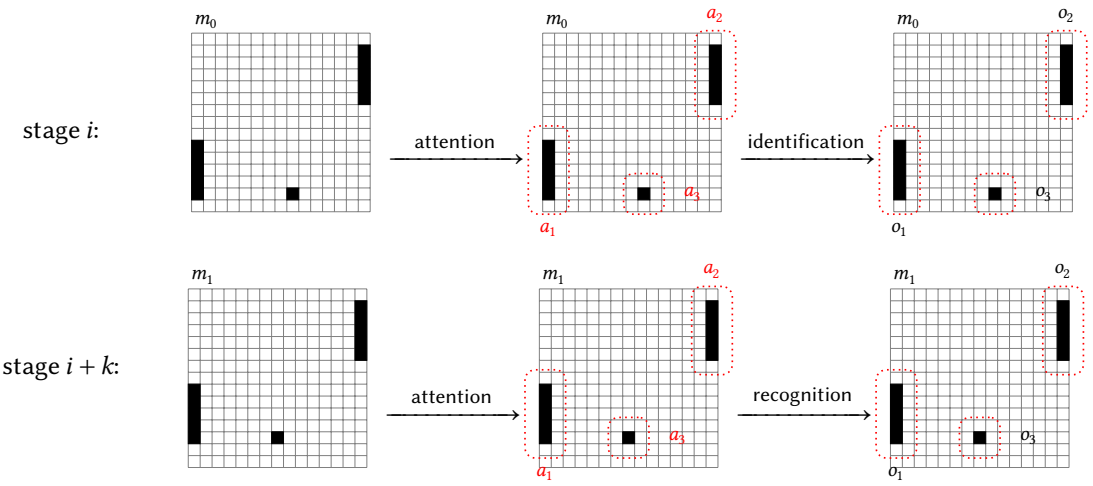

Figure 1: Initial image at subsymbolic level showing the identification of areas of interest in one stage, and the subsequent recognition of objects at a later stage.

2. Learning about Objects

Suppose we want to describe in a symbolic way a scene perceived as a sequence M of still images, each represented as a *binary* matrix m_i of $r \times c$ picture elements (pixels) $M = m_0, m_1, m_2, \ldots$. For simplicity, let us assume that these elements are binary, i.e., 0 or 1 (resp., 'white' – background, and 'black' – foreground). We can interpret each matrix as a snapshot of the world at a particular point in time and the sequence of matrices as the world's evolution over time. This paper describes what is involved, aiming to provide a blueprint for investigation.

In the first stage, we distinguish *objects* in the scene and associate *symbols* to them (an ongoing problem both in philosophy [3, 4, 5, 6] and computer science [7, 8]). While there have been prior formalisations of this problem [9, 10, 11], the notion of 'object' in our approach is central. Here we need to make suitable assumptions about what is of interest (to an agent) in the image and, for simplicity, assume that the black pixels are associated with objects of interest in the real world.

Consider the matrices in Figure 1. It is not difficult to identify at the subsymbolic (pixel) level the areas in red representing structured elements (or patterns), which we refer to as *objects*, within our unstructured input space. Let us assume that this can be done in an unsupervised manner using, e.g., an *attention function* [12, 13], defining areas of interest (AoIs) $a_1, a_2, a_3, \ldots$. This process should allow not only the creation of tracking mechanisms for the areas of interest in the images, but also the creation of new symbols for real-world objects and hence the establishment of an association between the real-world object, its perception and identification/recognition by the system (through some appropriate mechanism), and a symbolic representation. We aim to develop a framework in which objects, their properties, and the way in which they change can all be expressed and linked to these symbolic representations.

Our first objective is therefore to create a set of symbols and link these with a subsymbolic representation. To this end, the representation of an object, which we call a *signature*, exists in a more abstract space [14] than the direct visual representation of an AoI. The process of creating *new* symbols, e.g. o_1, o_2, o_3, and linking them to an abstract representation (see Figure 1, top), which we call *identification*, aims in part to allow these objects to be recognised in subsequent images. This is not trivial as differences in AoIs arise due to movement, inaccurate sensor information, etc., so there must be sufficient information to allow the association of the signature of an object in a new image with those from old images and the consequent retrieval of the previously associated symbols — we call this process *recognition* (see Figure 1, bottom).

By comparing distinct objects (in the same image) we can learn about similarities and differences [15, 16], generating a set of subsymbolic abstract *properties* ρ_1, ρ_2, ..., each assigned its own symbol, p_1, p_2, ..., thus linking the property's symbol to its subsymbolic representation in a process analogous to that of object-symbol creation. Initially, these properties are broad, yet object-specific — they are learnt from only one object. Once multiple objects have been identified with primitive properties created, these can be compared, finding similarities and distinctions that can be expressed as new properties themselves.

For example, comparing multiple objects that share a common feature, such as having the colour blue, etc, produces a set of similar 'shared properties'. If we apply the comparison again to the subsymbolic signatures of these shared properties, we can in turn figure out what they have in common, and what makes them distinct, eventually isolating a specific property of interest, such as 'blueness'. Once we have these primitives, we can re-define objects in terms of the properties they have. This has two important effects: it not only increases the vocabulary of our symbolic language, but it also allows objects to be symbolically represented in a non-atomic way. Non-atomicity is essential for the symbolic description of how an object changes (see Section 3). Of course, this suggests that the transition from subsymbolic to symbolic should be made through (several) layers of abstraction, until the desired level of granularity at the symbolic level can be achieved.

3. Reasoning About Transformations

Assuming we have mechanisms for uniquely identifying and recognising objects, we want to represent how objects change over time, aiming to express temporal transformations symbolically, retaining the association between the transformation and the representation. To be clear, if we perceive a change in the AoIs associated with a recognised object in successive images, we also want to associate a symbol with this change, allowing future reasoning about the transformation's 'meaning'.

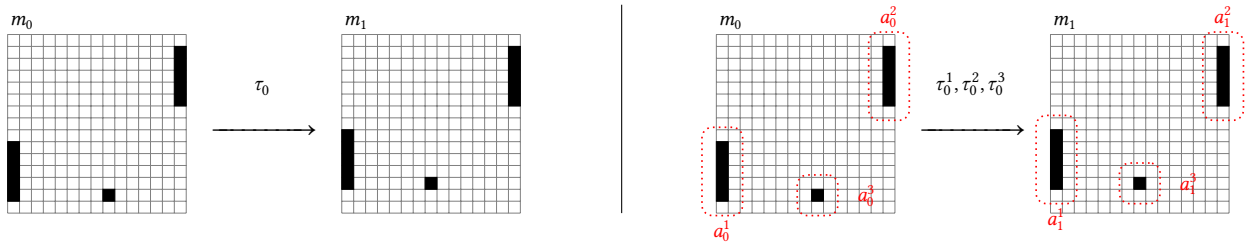

Figure 2: Variations between subsequent images, due to the leftmost paddle and ball moving up.

Consider the transformation τ_0 of matrix m_0 into m_1 on the left of Figure 2. At the subsymbolic level, we can consider the transformation applied to m_0 as a whole, but to reason about the *objects* in the image, we are more interested in the transformations applied to the AoIs a_1, a_2 and a_3, denoted by τ_0^1, τ_0^2, and τ_0^3, respectively (see right hand side of Figure 2). Assume that the area a_t^i is associated with object o_i in matrix m_t. For example, a_0^1 is associated with object o_1 in m_0 and a_1^1 with object o_1 in m_1, etc. Our task is to define the commutative diagram in Figure 3, so that $\sigma(a_t^i)$ is indeed a faithful representation of object o_i in m_t.

The $*$ in the commutative diagram of Figure 3, which shows the relationship between subsymbolic and symbolic levels, is intentional because in general multiple translations may be necessary to achieve the right level of abstraction at the symbolic level. In addition, the dotted lines indicate that the transformations may not necessarily be precise (see Section 4). This potential need for multiple translations is especially important in the generalisation of transformations involving the same object through a sequence of matrices.

Consider the problem of defining the concept of moving an object 'up' in a matrix and then generalising this notion to other objects. At the subsymbolic level, this operation transforms the region a_0^1 (of m_0) into a_1^1 (of m_1). We could give this transformation a symbol so that, e.g., the transformation τ_0^1 is represented by the *atomic* symbol $\sigma(\tau_0^1)$, but we would not be able to describe (in symbolic terms) what this transformation *entails*, hence it will be impossible to generalise it or to reason about it in more abstract terms. This means

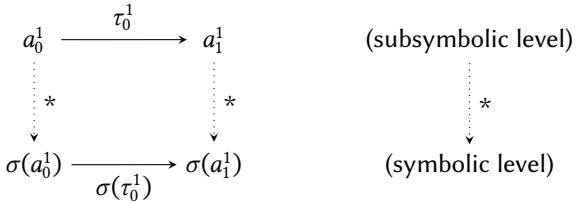

Figure 3: Relationship between subsymbolic and symbolic levels.

that in general we want $\sigma(a_0^1)$ *not* to be atomic in our symbolic language and, consequently, $\sigma(\tau_0^1)$ should be expressed in terms of how it affects some specific properties of objects (in this case o_1's 'location'). $\sigma(\tau_0^1)$ must provide a *faithful* representation of τ_0^1, meaning that the transformation from $\sigma(a_0^1)$ into $\sigma(a_1^1)$ must be such that $\sigma(a_1^1)$ indeed represents a_1^1, i.e., effectively how o_1 would be identified/recognised in m_1 (hence commuting the diagram of Figure 3). Because of the complexity of the operations involved, it may be sufficient for this process to simply be a "good enough" approximation.

The next question is then what should the appropriate level of granularity of these representations be for our exercise? The answer to this question is not simple, but in principle, we want the granularity to reflect the level of reasoning we hope to be able to capture. As an example, suppose that we need the relative positioning of the "boundaries" of o_1 within m_0 and m_1 (a reasonable assumption if we need to express the "movement" of objects). This means that we need to assume some coordinate/geometrical primitives which are available to us at the subsymbolic level, so that we can understand how to represent them symbolically. Thus, the identification process of o_1 must include these elements yielding a *signature* for o_1 that is not only sufficient for the recognition of o_1 in different matrices, but that also contains the basic ingredients for the description of the *properties* of o_1 that we need at the symbolic level. One such property could be, for example, o_1's *relative position* with respect to a common coordinate or to another area of interest (e.g., another object). Once we are happy with the basic components of o_1's signature we can associate it with a_0^1 and our job will be complete if $\sigma(\tau_0^1)$ is capable of producing a faithful symbolic representation of the signature a_1^1 in the form of $\sigma(a_1^1)$.

We gloss over a number of complications for now, but it should be easy to understand that we can only describe concepts at the symbolic level that we can somehow capture at the subsymbolic one. In practice, this means that what we associate with o_1 in Figure 1 at the subsymbolic level needs to capture enough of the relationship of o_1 with the rest of m_i so that we can describe it at the symbolic level in sufficient detail for the application in mind.

4. General Knowledge, Verification, Predictions and Revisions

Consider the game of Pong mentioned in Section 1, which is represented as a sequence of matrices $M = m_0, m_1, \ldots$ at the subsymbolic level. Some areas in the matrices are associated with objects, i.e., the ball and paddles. The objects may change in the sequence, e.g., by appearing in different locations in the matrices. These changes represent the notion of movement of the objects and can be used to describe how the scene evolves through time. We can perceive the changes between the matrices as transformations τ_0, τ_1, etc (see Figure 4). The challenge is to find a mechanism that describes the transformations in a way that is adequate for the application at hand. For example, we would like to describe τ_0 as a change in the relative y-coordinate of the left paddle and of the x-coordinate of the ball in m_1 with respect to their values in m_0 while still being able to associate the corresponding AoIs in m_1 with the 'same' objects identified in m_0. Eventually, we want to produce a sequence of representations $\sigma(a_0^i)$, $\sigma(a_1^i)$, etc, that describe the objects in a symbolic language, leading to symbolic representations $\sigma(m_0)$, $\sigma(m_1)$, ..., etc, of the matrices themselves.

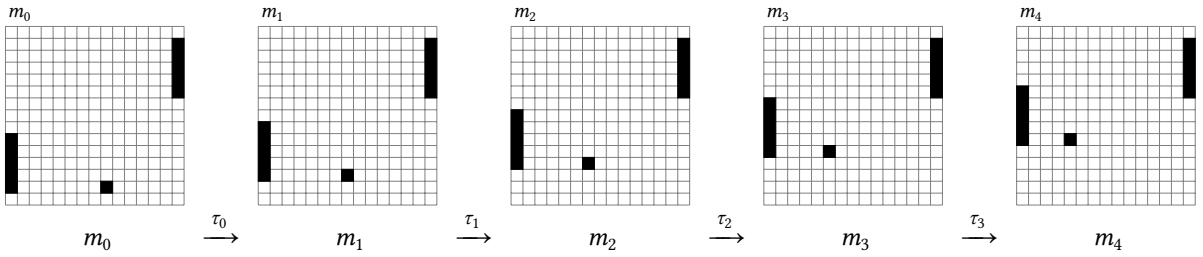

Figure 4: Transformations across a sequence of images.

To a large extent, we have only considered objects in isolation, but in general the subsymbolic representation may encode domain information about the objects' relationships that is valuable for symbolic reasoning. In our example, information embedded in the matrices m_i, such as the total number of objects, or their direction of travel with respect to each other, will not generally be captured by individual signatures of

objects identified. An important consideration is therefore how to incorporate general knowledge Γ that we need to add to $\sigma(m_j)$ s.t. $\Gamma \cup \sigma(m_j)$ is also faithful with respect to m_j for the particular application. Note that this notion of *general knowledge* is related (but distinct) to that of *common sense priors* mentioned in [17].

The transformations τ_i should eventually allow us to perform basic predictions about what future matrices should look like, e.g., through simulating the application of transformations to generate the next state. Verification of the predictions against actual inputs should provide opportunities for revision that generate more accurate translations with time. This results in a temporal aspect of the whole process as well. This dimension is depicted along the horizontal axis of the diagram of Figure 5.

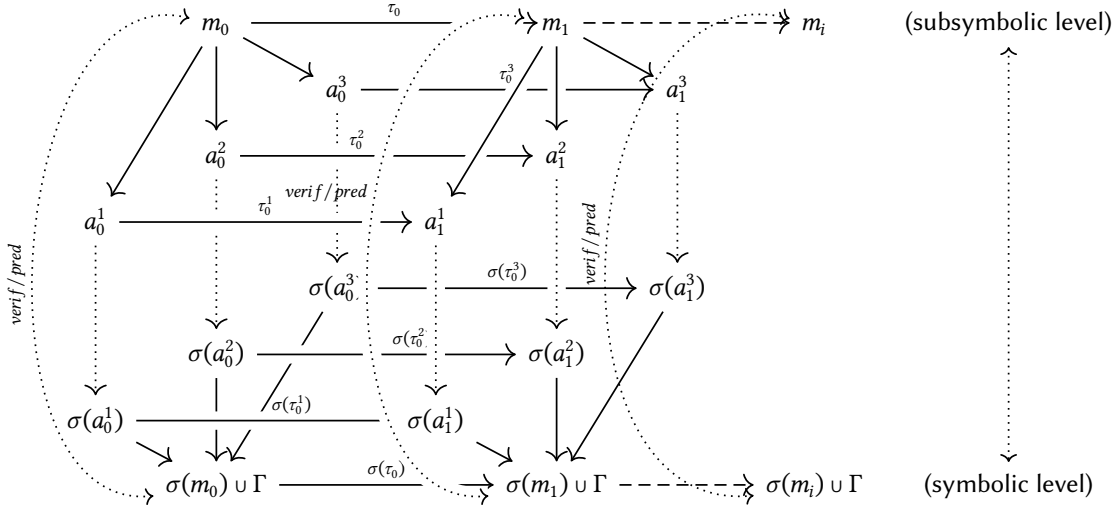

Figure 5: Relationship between subsymbolic and symbolic levels and successive matrices.

5. Conclusions

In this paper, we proposed the conceptualisation of a multi-layered framework for neuro-symbolic integration with learning. Although we have not provided a concrete instantiation, this level of separation between symbolic and sub-symbolic reasoning, and the proposed mode of integration, is novel and generalises other approaches by providing an initial scaffold upon which to employ specific techniques.

The initial focus is on how to define the basic ingredients with which a generic neuro-symbolic process can be devised that allows the association of subsymbolic components to symbolic counterparts, in a way that avoids hand-crafted symbols. We envisage the definition of object 'signatures' associating areas of interest in an image, the process that recognises them, and other components derived over time, and we postulate that comparisons between signatures can allow for the definition of some primitive concepts, such as properties, transformations, etc. More complex concepts can then be built from these in layers at increasing levels of abstraction, eventually leading to the development of a symbolic language over time arising from completely subsymbolic inputs.

Finally, we see revisions of the model arising through comparisons of predictions of future states with actual inputs. In future work, we will investigate how interactions and relationships *between* objects can be captured symbolically using these ideas.

Acknowledgments

This work was supported by UK Research and Innovation grant number EP/S023356/1, in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (https://www.safeandtrustedai.org).

References

[1] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people, Behavioral and Brain Sciences 40 (2017) e253. doi:10.1017/S0140525X16001837.

[2] B. Goertzel, Perception processing for general intelligence: Bridging the Symbolic/Subsymbolic gap, in: J. Bach, B. Goertzel, M. Iklé (Eds.), Artificial general intelligence, volume 7716 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 79–88. doi:10.1007/978-3-642-35506-6_9.

[3] M. Taddeo, L. Floridi, Solving the Symbol Grounding Problem: A Critical Review of Fifteen Years of Research, Journal of Experimental & Theoretical Artificial Intelligence 17 (2005) 419–445. doi:10.1080/09528130500284053.

[4] A. Cangelosi, A. Greco, S. Harnad, Symbol Grounding and the Symbolic Theft Hypothesis, in: A. Cangelosi, D. Parisi (Eds.), Simulating the Evolution of Language, Springer, London, 2002, pp. 191–210. URL: https://doi.org/10.1007/978-1-4471-0663-0_9. doi:10.1007/978-1-4471-0663-0_9.

[5] L. Steels, The symbol grounding problem has been solved. So what's next?, in: M. de Vega (Ed.), Symbols and embodiment: Debates on meaning and cognition, Oxford University Press, Oxford, 2008.

[6] S. Bringsjord, The symbol grounding problem … remains unsolved, Journal of Experimental & Theoretical Artificial Intelligence 27 (2015) 63–72. doi:10.1080/0952813X.2014.940139.

[7] R. Cubek, W. Ertel, G. Palm, A Critical Review on the Symbol Grounding Problem as an Issue of Autonomous Agents, in: S. Hölldobler, R. Peñaloza, S. Rudolph (Eds.), KI 2015: Advances in Artificial Intelligence, volume 9324 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2015, pp. 256–263. doi:10.1007/978-3-319-24489-1_21.

[8] S. Coradeschi, A. Loutfi, B. Wrede, A Short Review of Symbol Grounding in Robotic and Intelligent Systems, KI - Künstliche Intelligenz 27 (2013) 129–136. doi:10.1007/s13218-013-0247-2.

[9] K. Greff, S. van Steenkiste, J. Schmidhuber, On the Binding Problem in Artificial Neural Networks, arXiv (2020). URL: http://arxiv.org/abs/2012.05208.

[10] T. R. Besold, K.-U. Kühnberger, A. S. d. Garcez, A. Saffiotti, M. H. Fischer, A. Bundy, Anchoring knowledge in interaction: Towards a harmonic Subsymbolic/Symbolic framework and architecture of computational cognition, in: J. Bieger, B. Goertzel, A. Potapov (Eds.), Artificial general intelligence, volume 9205 of *Lecture Notes in Computer Science*, 2015, pp. 35–45. doi:10.1007/978-3-319-21365-1_4.

[11] T. R. Besold, A. S. d. Garcez, S. Bader, H. Bowman, P. M. Domingos, P. Hitzler, K.-U. Kühnberger, L. C. Lamb, D. Lowd, P. M. V. Lima, L. de Penning, G. Pinkas, H. Poon, G. Zaverucha, Neural-symbolic learning and reasoning: A survey and interpretation, arXiv (2017). URL: http://arxiv.org/abs/1711.03902.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017, pp. 5998–6008.

[13] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, A. Lerchner, MONet: Unsupervised Scene Decomposition and Representation, arXiv (2019). URL: http://arxiv.org/abs/1901.11390.

[14] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, J. Wu, The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision, in: 7th international conference on learning representations, ICLR 2019, 2019. URL: https://openreview.net/forum?id=rJgMlhRctm.

[15] P. J. Blazek, M. M. Lin, A neural network model of perception and reasoning, arXiv (2020). URL: http://arxiv.org/abs/2002.11319.

[16] L. A. A. Doumas, G. Puebla, A. E. Martin, J. E. Hummel, Relation learning in a neurocomputational architecture supports cross-domain transfer, in: S. Denison, M. Mack, Y. Xu, B. C. Armstrong (Eds.), Proceedings of the 42nd annual meeting of the cognitive science society, CogSci 2020, Montreal, 2020, pp. 932–937. URL: https://cogsci.mindmodeling.org/2020/papers/0165/index.html.

[17] M. Garnelo, K. Arulkumaran, M. Shanahan, Towards Deep Symbolic Reinforcement Learning, arXiv (2016). URL: http://arxiv.org/abs/1609.05518.