Towards A Dual Process Approach to Computational Explanation in Human-Robot Social Interaction

Agnese Augello, Ignazio Infantino, Antonio Lieto*, Umberto Maniscalco, Giovanni Pilato, Filippo Vella
Institute of High Performance Computing and Networking,
National Research Council, ICAR-CNR, Palermo, Italy
name.surname@cnr.it, lieto@di.unito.it
*Dipartimento di Informatica, University of Turin, Italy

Abstract

The capacity for AI systems of explaining their decisions represents nowadays a huge challenge for both academia and industry (e.g. let us think at the autonomous cars sector). In this paper we sketch a preliminary proposal suggesting the adoption of a dual process approach for computational explanation. Our proposal is declined in the field of Human-Robot Social Interaction; namely, in a gesture recognition task.

1 Introduction

The capability for AI systems of providing an explanation about the reasons guiding their decisions, represents a crucial challenge and research objective in the current fields of Artificial Intelligence (AI) and Computational Cognitive Science [Langley *et al.*, 2017]. Current AI systems, in fact, despite the enormous progresses reached in specific fields, mostly fail to provide a transparent account of the reasons determining their behavior (both in cases of a successful or unsuccessful output). This is due to the fact that the adoption of current Machine Learning and Deep Learning techniques faces the classical problem of opacity in neural networks; furthermore, this problem explodes with the current techniques[1]. In our opinion a possible way to deal with this problem is based on a dual process approach.

The dual process theories of mind [Evans and Frankish, 2009; Stanovich and West, 2000; Kahneman, 2011] is an experimentally grounded theory proposed in the field of psychology of reasoning. It suggests that our cognition is governed by two types of interacting cognitive systems, which are called respectively system(s) 1 and system(s) 2. Systems of the type 1, referred also as S1, operate with rapid, automatic, associative processes of reasoning. They are phylogenetically older and execute processes in a parallel and fast way. Type 2 systems, referred also as S2, are, on the other hand, phylogenetically more recent and they are based on conscious, controlled, sequential processes (also called type 2 processes) and on logic-based rule following. As a consequence, if compared to system 1, system 2 processes are slower and cognitively more demanding. In the dual process perspective, then, decision making consists in a two-step procedure based on the interaction between heuristic, perception-guided (and biased) thinking (type 1 processes), with forms of deliberative thinking based on the canons of normative rationality (and on type 2 processes). In recent years, the cognitive modeling and the AI community have posed a growing attention on the dual process theories as a framework for modeling artificial cognition. Efforts have been made, for example, in the areas of knowledge representation and reasoning [Frixione and Lieto, 2014], cognitive systems dealing with arithmetical calculations [Strannegård *et al.*, 2013], cognitive models of emotions [Larue *et al.*, 2012], question answering for common-sense linguistic descriptions ([Lieto *et al.*, 2015], [Lieto *et al.*, 2017b]), computational creativity [Augello *et al.*, 2016b] as well as in the design of general purpose cognitive architectures, such as CLARION, whose principles are explicitly inspired by such a theoretical framework [Sun, 2006].

In this position paper, we propose to consider a dual process approach to deal with the problem of computational explanation (or, at least, with one facet of this problem). In doing so we propose to endow a social robot with a computational explanation module based on different components: a S1 and a S2 one. The S1 module is responsible for the fast categorization and for the perceptual based recognition of gestures in a social context (and is based on deep neural network architecture) while the S2 component is responsible for providing a high level model that can be exploited to extract an explanation about the 'reasons', e.g. the high level features, that characterize the categorized output provided by S1[2]. Such descriptive model, based on explicit representations, can be used to formulate plausible forms of explanation about the features/properties that are usually considered to lead to the particular classification of gestures provided by the, opaque, S1 component. In particular, S2 exploits an ontology that describes the features characterizing the follow-

[1] Despite some proposals exists in the literature reporting sparse cases where a partial interpretation of the operations of the units is possible [Zhou *et al.*, 2015] as well as alternative proposals are available in order to reduce the opacity in neural networks[Lieto *et al.*, 2017a], the general problem still remains unsolved

[2] Currently the S2 component is activated on demand. On the difficulty concerning the question about 'when' the S2 process is activated we refer to [Lieto *et al.*, 2018]).

ing actions: “Bowing”, “Clapping”, “Handshaking”, “Punching”, “Slapping” and “Frontkicking”. The first three of them are grouped into a “meta-class” of *normal* actions, while the remaining three are grouped as *aggressive* ones. Once the S1 module gives its output, the S2 module acquires the S1 outcome and tries to explain, if required by the user, what characterizes the perceived action according to the knowledge coded into the ontology.
In the near future the authors will implement the proposed approach in a Pepper robot, improving also the interaction between S1 and S2. In the next sections we provide some additional details of the proposed framework.

2 The Framework

Figure 1 shows a simplified schema of the proposed framework. An Artificial Intelligence system (AI Module) that drives robot social activities is responsible for the perception, processing, and action of the social robot (S1). The perception capabilities of the robot have to detect relevant features of human social behavior.
RGB cameras, microphones, lasers, sonars, depth cameras, and other sensors allow the robot to capture different aspects of the human action and the context. Data arising from each sensor require complex computation and a different level of abstraction to produce an appropriate robot reaction. The observed robot behavior, in part designed by programmers, could be difficult to understand to the final user because it depends on how the robot interprets the real perception.
As reported in Figure 1, many social relevant entities should be processed to determine a realistic robot social behavior: speech, facial expressions, sounds, but also environment features that influence the social context (see for example [Infantino *et al.*, 2008], [Infantino *et al.*, 2007]).
The robot typically interacts with the human by a verbal output and postures (Animated Say). The social interaction is subject to both an internal and external evaluations. The internal assessment considers the robot aim that consists of detecting a desired human state. The external evaluation is directly given by the human user involved in the interaction or by an observer. Furthermore, if required, the robot should enable the computational explanation process that justifies its verbal outputs and actions.
In this proposal we focus the attention on a single perceptual input, given by an RGBD camera.This device is used to catch human social signs occurring during the interaction between the robot and an human being.
The infrared laser of the RGBD camera and its receiver detect a set of 3D points, from which the human skeleton is extracted. The analysis of the temporal evolution of skeleton joints allows the system to detect relevant action (i.e. a couple constituted by an initial posture and a final one). A deep neural network (DNN) approach tries to classify the social sign. Such machine learning methodology requires a training using a given dataset of postures. Naturally, the learning phase determines the interpretation of the detected social signs, and the subsequent computational processes to decide robot reactions.
The reasoning, execution and evaluation module together with perceptive capabilities is organized following a dual process theory paradigm (see for example [Augello *et al.*, 2016b]) that, in the present setting, has been extended and used to enable a computational explanation subsystem.

3 Social interaction scenario

During a social interaction, the involved agents usually follow “social practices” [Reckwitz, 2002], i.e. routinized behaviors depending on the social context, the mutual expectations, and the pursued aims. A “social intelligent” robot must be able to properly manage these practices, understanding the social situation and consequently planning and adapting its behavior [Dignum *et al.*, 2014] [Augello *et al.*, 2016a]. A key role in the understanding of the situation is given by the interpretation of all the possible social signs expressed, both in a verbal and not verbal ways, by the agents involved in the interaction. The interpretation of these signals, and in particular the not verbal ones, allows an agent to recognize and understand the intentions, emotions and attitudes of people.
Let us focus on a practice of reception in a public office, considering the task of welcoming visitors in the waiting room and directing them to proper office rooms.
In this scenario, the robot must be able to discriminate the not appropriate behaviors of the visitors. For example, someone can become unsettled if he must wait too much in the waiting room or if there is some problem with the appointment. The robot learns how to detect not appropriate and in particular aggressive behaviors, by examining the postures and the gestures of people during a training phase. Then, during the interaction, considering its expectations and its experience, he must be able to quickly recognize the exhibited social signs. Then, if required, he must be able to provide an explanatory account of some sort of this process of interpretation.
Classification and explanation are respectively accomplished through S1 and S2 processes and components as briefly described in the following sections.

3.1 Social Signs Interpreation Process

The robot is capable to classify the social signs represented through 3D human postures by using a Deep Neural Network approach. An RGBD device captures the human skeleton as a spatial localisation of the relevant joints.

In the last years the techniques of deep learning have given a strong impulse for the machine learning algorithms. New achievements have been shown in the generalization capabilities of the input pattern and the discrimination of different classes. These networks can learn to discriminate a big set of data with their labels.

The great advantage of the deep networks is that the first layers of the network, if suitably trained, can automatically extract features allowing a robust representation of the input pattern. In this way, instead of using hand crafted features, the raw data can be processed creating a good classifier with features learned from data [LeCun *et al.*, 2015]. Beyond the possibility of extracting spatial features for the recognition of patterns in bidimensional inputs, deep networks can also be used for the processing and classification of sequence of

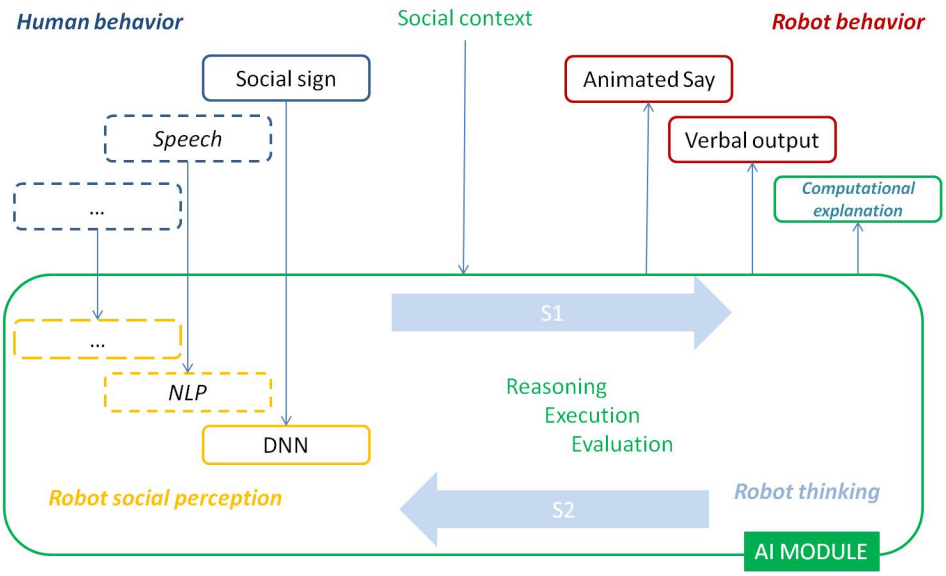

Figure 1: The computational explanation framework for the social robot scenario

data. In these cases, the network, through recurrent connections can maintain memory of the past history and compute the input accordingly. An example of these kind of networks are the Long Short Term Memory networks that are used in the proposed 2 system.

LSTMs have been designed by Hochreiter and Schmidhuber [Hochreiter and Schmidhuber, 1997] with the aim of avoiding the long-term dependency problem, at the price of a more complex cell structure. The key feature of LSTMs is the "cell state" that is propagated from a cell to another. State modifications are regulated by three structures called gates, composed out of a sigmoid neural net layer and a pointwise multiplication operation. Let $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$ the cell state at time t; the first gate, called "forget gate layer", considers both the input x_t and the output from the previous step h_{t-1}, and returns values between 0 and 1, describing how much of each component of the old cell state C_{t-1}, where should be left unaltered: if the output is 0, no modification is made; if the output is one, the component is completely replaced. New information to be stored in the state is processed afterwards. The second sigmoid layer, called the input gate layer decides which values will be updated. Next, a $tanh$ layer creates a vector of new candidate values, $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_c)$, that could be added to the state. To perform a state update, C_{t-1} is first multiplied by the output of the forget gate $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$, and the result is added to the pointwise multiplication of the input gate output $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$ and $\tilde{C}_t$. Finally, the output $h_t = o_t * \tanh(C_t)$ can be generated, where $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$. First, a sigmoid is applied, taking into account both h_{t-1} and x_t; its output is then multiplied by a constrained version of C_t, so that we only output the parts we decided to. In [Pascanu *et al.*, 2013] it is given a detailed theoretical explanation of the reasons behind the advantages of using a network made of multiple layers. In our scenario we have chosen to gradually stack LSTM layers and measure the trend of the F1-score to determine what the correct number of layers can be. Each LSTM layer is separated from the next one by a ReLU function. In addition, given a sequence length, we attempted to determine how many neurons are needed for the representation to be of good quality. To speed up the information acquisition task, to train the network, has been used a dataset formed by a set of actions divided in two macro classes dealing with aggressive or non aggressive behavior [Theodoridis and Hu, 2007]. The dataset has been created by monitoring, by using proper sensors, placed in several body parts, the free movements of ten people in front of a camera or in front of a standing bag. The actions of the dataset, with twenty different labels, are divided between actions for the *normal* behavior and actions for *"not friendly"* behavior.

3.2 The social sign dataset

For the experiments we have used a subset of the Vicon Physical Action dataset first used in [Theodoridis and Hu, 2007] and made available through [Lichman, 2013].
10 subjects (7 men and 3 women) have been recorded while performing 20 actions, each accounting for 10 normal and 10 not friendly activities. For our setup, a subset of the actions, more inherent to the considered context, has been selected, composed of three normal actions (Bowing, Clapping and Handshaking) and three aggressive actions (Punching, Slapping and Frontkicking). The training has been performed on nine subjects, while testing is done on the tenth, last subject. We have tested with the following variations: the number of neurons in the LSTM layers have been set to the values 64, 128 or 256; one, two or three stacked LSTM levels have been considered and the sliding window have been set to a value from 2 to 20. The training has been performed for 10 epochs.

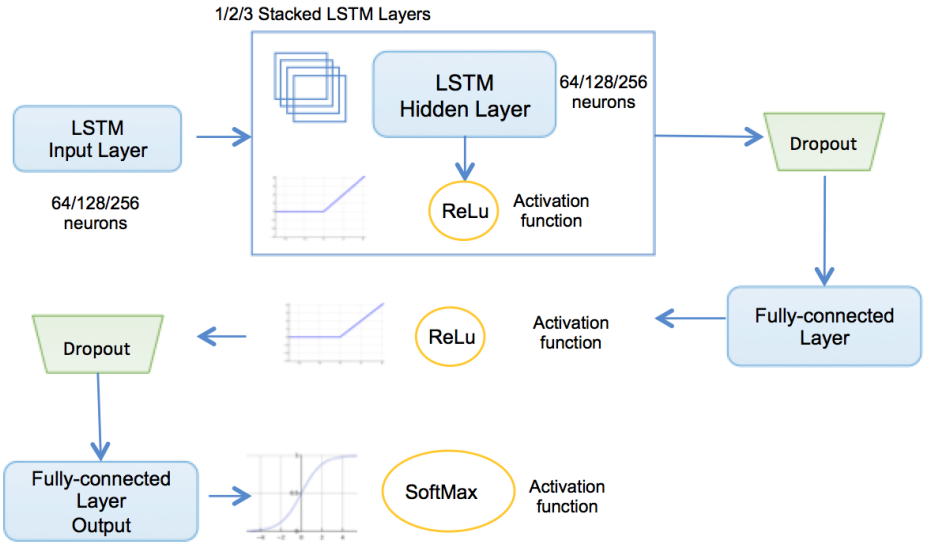

Figure 2: Proposed network architecture

After 10 epochs, the accuracy has already approached 1, so we choose to stop.

3.3 A Partial Explanatory Account via Ontology

In the current, preliminary, version of our system, the S2 component is based on an a general ontological model representing the main perceptual differences between different classes of gestures (e.g. aggressive vs not aggressive ones). The representational differences between the two ontological states is exploited to provide an high level description about the reasons leading to categorize a particular perceived gesture as being aggressive or not. The ontological domain features considered to distinguish among these two classes are, for example, velocity of the gesture execution (usually aggressive behavior proceeds in a fast way), distance of the final gesture position from the body etc[3]. The current version of the ontology is available, also in a navigable format, at `http://www.di.unito.it/~lieto/ExpActOnto.html` In addition, the S2 component also takes into account the analysis of the training of the S1 neural component (given a specific dataset). In particular, since the Neural Network used in S1 is capable of categorizing a gesture into a set of 20 kinds of actions, while the current ontology is capable to describe only 6 among them, we compute the similarity/dissimilarity of the outcome given by S1 with respect to the six actions modeled in S2. This allows us to attribute the unknown perceived gesture to a known one, following a typical Case-Based Reasoning approach.
The S2 component allows also to model the differences between gestures of the same meta-level classes (i.e. the aggressive and normal ones). These sub-models allow to represent in more detail the differences of similar gestures and can be used to described why a particular sign, e.g. firstly categorized as 'aggressive', has been additionally recognized, for example, as a 'Punching'. Figure 3 shows an example of this kind of explanation. Namely: it shows that a 'Punching' Action is characterized by the fact of being an action executed at a certain velocity (X), categorized as 'High Velocity', and at a certain distance (Y) from the Body, categorized as 'Close Distance' according to the ontology. In addition to these traits, common to all the 'Aggressive Actions', the 'Punching' action is also characterized by the fact of being executed with "Close Hands". Figure 4, finally, provides an additional model-based explanation about why the previous 'Punching' cannot be classified, for example, as a 'Slapping' (both 'Slapping' and 'Punching' are 'Aggressive Actions'). Also in this case the fact that the detected body part executing the gesture is a 'Close Hand' and not a 'Open Hand' (as in the case of 'Slapping') represent a crucial element for explaining the reason leading to that categorization decision.

[3]In other words: we try to provide an explanatory account of the output of the opaque S1 component by using an *apriori* ontological model of a given situation.

4 Conclusions and Future Work

In this paper we have sketched a preliminary account of a dual process based framework able to provide a partial explanation of the reasons driving a robotic system to some decisions in task of gesture recognition is a social scenario. In the near future we plan to evaluate in detail the feasibility of the proposed framework with a Pepper robot. In addition, as mid term goal, we plan to extend the level of detail of the possible explanation provided by such framework by considering more complex scenarios and a multimodal interaction involving both visual and linguistic elements. Finally, we plan to provide a tighter integration of the two software components that, currently, operate in a relatively independent way.

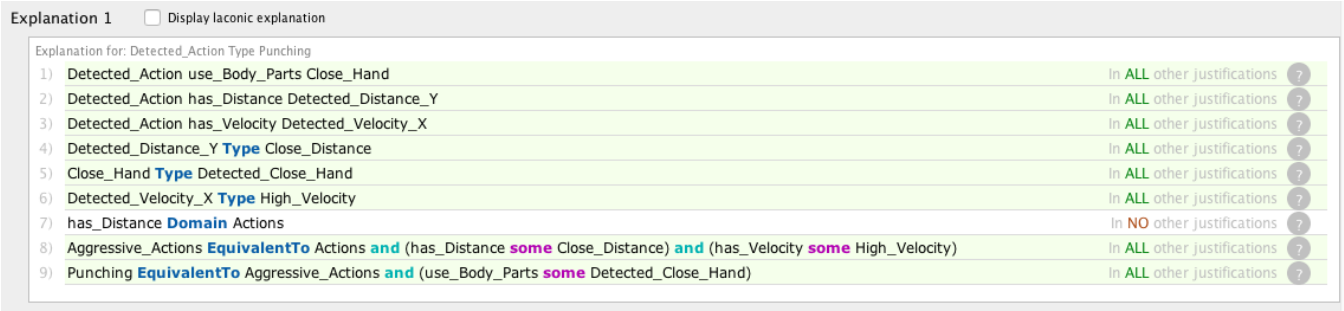

Figure 3: Action Explanation for Punching

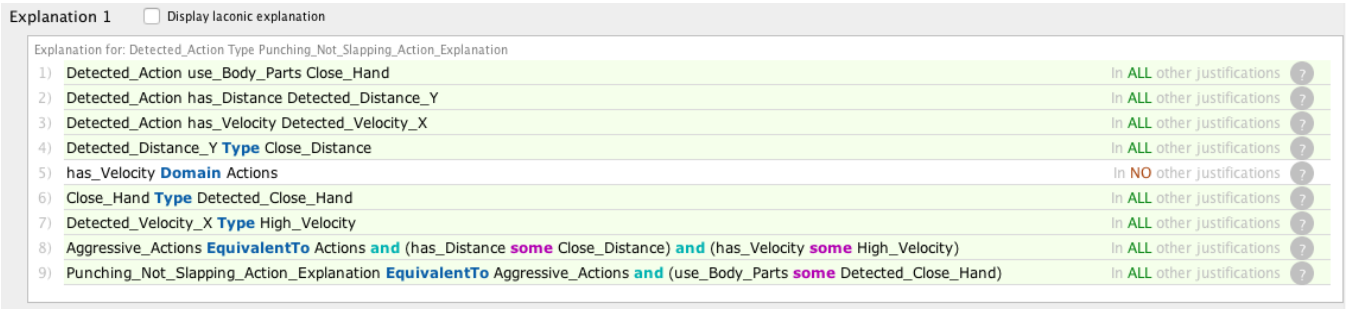

Figure 4: Explaining why Punching and not Slapping

References

[Augello *et al.*, 2016a] Agnese Augello, Manuel Gentile, and Frank Dignum. Social agents for learning in virtual environments. In *Games and Learning Alliance*, pages 133–143. Springer, 2016.

[Augello *et al.*, 2016b] Agnese Augello, Ignazio Infantino, Antonio Lieto, Giovanni Pilato, Riccardo Rizzo, and Filippo Vella. Artwork creation by a cognitive architecture integrating computational creativity and dual process approaches. *Biologically Inspired Cognitive Architectures*, 15:74–86, 2016.

[Dignum *et al.*, 2014] Virginia Dignum, Catholijn Jonker, Rui Prada, and Frank Dignum. Situational deliberation; getting to social intelligence. *Computational Social Science and Social Computer Science: Two Sides of the Same Coin*, 2014.

[Evans and Frankish, 2009] Jonathan St BT Evans and Keith Ed Frankish. *In two minds: Dual processes and beyond.* Oxford University Press, 2009.

[Frixione and Lieto, 2014] Marcello Frixione and Antonio Lieto. Towards an Extended Model of Conceptual Representations in Formal Ontologies: A Typicality-Based Proposal. *Journal of Universal Computer Science*, 20(3):257–276, March 2014.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Infantino *et al.*, 2007] Ignazio Infantino, Riccardo Rizzo, and Salvatore Gaglio. A framework for sign language sentence recognition by commonsense context. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(5):1034–1039, 2007.

[Infantino *et al.*, 2008] Ignazio Infantino, Carmelo Lodato, Salvatore Lopes, and Filippo Vella. Human-humanoid interaction by an intentional system. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 573–578. IEEE, 2008.

[Kahneman, 2011] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[Langley *et al.*, 2017] Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. Explainable agency for intelligent autonomous systems. 2017.

[Larue *et al.*, 2012] Othalia Larue, Pierre Poirier, and Roger Nkambou. A cognitive architecture based on cognitive/neurological dual-system theories. In *International Conference on Brain Informatics*, pages 288–299. Springer, 2012.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[Lichman, 2013] M. Lichman. UCI machine learning repository, 2013.

[Lieto *et al.*, 2015] Antonio Lieto, Andrea Minieri, Alberto Piana, and Daniele P. Radicioni. A knowledge-based system for prototypical reasoning. *Connection Science*, 2015.

[Lieto *et al.*, 2017a] Antonio Lieto, Antonio Chella, and Marcello Frixione. Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation. *Biologically Inspired Cognitive Architectures*, 19:1–9, 2017.

[Lieto *et al.*, 2017b] Antonio Lieto, Daniele P Radicioni, and Valentina Rho. Dual peccs: a cognitive system for conceptual representation and categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2):433–452, 2017.

[Lieto *et al.*, 2018] Antonio Lieto, Christian Lebiere, and Alessandro Oltramari. The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, 48:39–55, 2018.

[Pascanu *et al.*, 2013] Razvan Pascanu, Guido Montúfar, and Yoshua Bengio. On the number of inference regions of deep feed forward networks with piece-wise linear activations. *CoRR*, abs/1312.6098, 2013.

[Reckwitz, 2002] Andreas Reckwitz. Toward a theory of social practices: A development in culturalist theorizing. *European journal of social theory*, 5(2):243–263, 2002.

[Stanovich and West, 2000] Keith E Stanovich and Richard F West. Advancing the rationality debate. *Behavioral and brain sciences*, 23(05):701–717, 2000.

[Strannegård *et al.*, 2013] Claes Strannegård, Rickard von Haugwitz, Johan Wessberg, and Christian Balkenius. A cognitive architecture based on dual process theory. In *International Conference on Artificial General Intelligence*, pages 140–149. Springer, 2013.

[Sun, 2006] Ron Sun. The CLARION cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and multi-agent interaction*, pages 79–99, 2006.

[Theodoridis and Hu, 2007] Theodoros Theodoridis and Huosheng Hu. Action classification of 3d human models using dynamic anns for mobile robot surveillance. In *Robotics and Biomimetics, 2007. ROBIO 2007. IEEE International Conference on*, pages 371–376. IEEE, 2007.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2015.