IEC
Institute of Electronics and Computer

Journal of Artificial Intelligence and Systems, 2022, 4, 83-106
https://iecscience.org/journals/AIS
ISSN Online: 2642-2859

Neural Network Learning of Context-Dependent Affordances

Luca Simione[1, a, *], Anna M Borghi[1, 2, b], Stefano Nolfi[1, c]

[1] Istituto di Scienze e Tecnologie della Cognizione, CNR, Rome, Italy
[2] Dipartimento di Psicologia Dinamica, Clinica e Salute, Sapienza, Università di Roma, Rome, Italy
[a]luca.simione@istc.cnr.it, [b]anna.borghi@uniroma1.it, [c]stefano.nolfi@istc.cnr.it
*Corresponding Author

How to cite this paper: Luca Simione, Anna M Borghi, Stefano Nolfi (2022). Neural Network Learning of Context-Dependent Affordances. Journal of Artificial Intelligence and Systems, 4, 83–106. https://doi.org/10.33969/AIS.2022040106.

Received: March 9, 2022
Accepted: November 22, 2022
Published: December 9, 2022

Copyright © 2022 by author(s) and Institute of Electronics and Computer. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

In this paper, we investigated whether affordances are activated automatically, independently of the context in which they are experienced, or not. The first hypothesis postulates that stimuli affording different actions in different contexts tend to activate all actions initially. The action appropriate to the current context is later selected through a competitive process. The second hypothesis instead postulates that only the action appropriate to the current context is activated. The apparent tension between these two alternative hypotheses constitutes an open issue since, in some cases, experimental evidence supports the context-independent hypothesis, while in other cases it supports the context-dependent hypothesis.

To study this issue, we trained a deep neural network with stimuli in which action inputs co-varied systematically with visual inputs. The neural network included two separate pathways for encoding visual and action inputs with two hidden layers each, and then a common hidden layer. The training was realized through an auto-associative unsupervised learning algorithm and the testing was conducted by presenting only part of the stimulus to the neural network, to study its generative properties.

As a result of the training process, the network formed visual-action affordances. Furthermore, we conducted the training process in different contexts in which the relation between stimuli and actions varied. The analysis of the obtained results indicates that the network displays both a context-dependent activation of affordances (i.e., the action appropriate to the current context tends to be more activated than the alternative action) and a competitive process that refines action selection (i.e., that increases the offset between the activation of the appropriate and unappropriate actions). Overall, this suggests that the apparent contradiction between the two hypotheses can be resolved. Moreover, our analysis indicates that the greater facility with which colour-action associations are acquired with respect to shape-action associations is because the representation of surface features, such as colour, tends to be more readily available for deeper features, such as shape.

Our results support the feasibility of human-like affordance acquisition in artificial

neural networks trained using a deep learning algorithm. This model could be further applied to a number of robotic and applicative scenarios.

Keywords

Deep learning, neural network, affordance, context, action

1. Introduction

The ability to respond adequately to the 'invitations' that objects in the environment offer us is crucial to the survival of our species. In recent years, an increasing number of studies have focused on affordances, i.e., on objects invitations to act. Since the seminal work of Gibson [1], affordances have become highly interesting for recent research in cognitive science, cognitive neuroscience, and robotics [2], also due to the spread of the embodied and grounded cognition view [3]–[7]. The idea underlying the notion of affordances, that observing objects leads to the activation of motor responses, is highly appealing for embodied and grounded cognition views, according to which perception, action, and cognition are strictly interrelated.

One of the most debated issues about affordances in the literature concerns their activation. The question is whether object affordances are automatically activated or only activated when relevant to the current context (i.e., to the current situation and/or goal). Most studies in the late ‘90s and early 2000 stressed the automaticity of affordances. Their goal was to demonstrate that affordances were activated automatically, independently of the context and task. Here is an example of a very influential experiment conducted by Tucker and Ellis [8]: participants were required to respond whether images of common objects (e.g., knife, pan) were upright or reversed by pressing two different keys on the keyboard. Responses were faster when the location of the object handle (left, right) and the location of the key to press to provide a response (left, right) corresponded [8]. This result suggests that participants were sensitive to the object's shape and the location of its parts, even if the task did not require paying attention to it. Similarly, another influential study showed that even if participants were required to respond to the object category (artefacts vs. natural objects), their motor responses took into account object size [9], [10].The developmental and neural evidence converges in highlighting the automaticity of affordances [11].

The view according to which affordances are automatically activated has recently been challenged. Recent research has in fact started to focus on the role context plays in affordance activation [11]. For example, behavioural research has shown that object affordances are activated when objects are located in the near but not in the far space [12], [13]. Furthermore, the presence of multiple visual objects and different

task requests modulate the affordance effect on response speed [14]. Recent studies have shown that, depending on the context, the affordances related to object manipulation or function are activated [15]–[17] and have revealed the influence of the social context on the activation of the affordances [18]–[22]. For example, the presence of others modulates the activation of affordances, and this modulation might differ depending on whether others have a positive or negative attitude. Brain imaging studies revealed that motor areas are active during manipulable object processing, but not function related objects [23].

Whether these two positions can come to a conciliation is currently an open issue. Further research concerning, for example, the time course of affordance activation is needed to better understand the mechanisms of affordance selection. One possibility is that the context selects only the relevant affordances. For example, only the affordances of a cup's handle would be recruited in the context of drinking. This possibility is incompatible with the idea that affordances are automatically activated. Another possibility is that all affordances are automatically activated, and the context acts as a sort of "late" filter, selecting only the relevant ones. This possibility is compatible with the idea of automaticity of affordances. Consistent with this view, an influential model shows that in the brain, different affordances and action possibilities could compete [24], [25]. If this is the case, all affordances of the cup, not just the handle, should be early activated; then, in a drinking context, only the handle should win the competition among affordances.

The debate we have illustrated motivates numerous recent studies. While until the early 2000s, most studies focused on the automaticity of affordances, recent literature addresses whether the characteristics of the task/context influence the activation of affordances. Particularly relevant to the present work are studies showing that affordance effects emerge solely when the task involves a deep level of processing – for example, when it requires processing object shape but not object colour. For example, Tipper, Paul, and Hayes [26] used a variation of the paradigm by Tucker and Ellis [8] to investigate the automatic compatibility effect found when participants process object shape, which is a property relevant to grasp affordances and object colour, which is not. They found clear affordance effects when the task required to process object shape but no effect when it required to process colour [27].

The reported studies suggest that context/task modulates affordance activation. However, they leave some important issues unsolved. It is indeed unclear why affordance effects were found during shape but not during colour processing. Some possibilities can be devised. The results can be due to either the lower complexity of color compared to shape or the link between shape and grasping actions (for similar conclusions with a different experimental paradigm, see [28]). Both hypotheses have

the consequence that colour is processed more superficially compared to shape, and both can therefore cause the absence of the affordance effects.

The present study aims to deepen the previously discussed issues by using neural networks trained to associate different objects' properties (first orientation, then colour and shape) to action. Performing a study with neural networks has many advantages. The most important in this context is that, unlike humans, neural networks are not biased by previous knowledge; therefore, it is possible to control how they learn from scratch a given association [29]. Our study has two main objectives. First, it aims to test whether the affordances are automatically activated, whether their activation is context-dependent, and whether the eventual activation of multiple affordances is resolved later through a competitive process [30], [31]. To test these different hypotheses, we trained a deep neural network to form affordances between a series of experienced objects and actions in two contexts in which the relation between the experienced object and the action depended on the object orientation or the object shape/colour, respectively. The second aim of the study is to investigate the role of the feature type (i.e., colour versus shape) in the formation of the new affordances and the resolution of the possible conflicts that might arise in the case of objects affording multiple context-dependent actions.

The overall objective of this work is to extend our understanding of the role of affordance in humans. We thus use neural network and deep learning methods to collect synthetic data which complement the data analyzed in experimental studies. For related works which aim to develop effective robots without necessarily modelling human behaviour, see [32], [33]. For a general review of research on affordance in robotics, see [34]. We aim to bridge experimental data collected with human or primate subjects with neural network models, in which experimental data could be replicated and neural mechanisms underpinning them could be further investigated. Primary, the novelty of our work lies in demonstrating how a deep neural network trained from scratch would be capable of replicating interesting and not yet fully understood human processes such as affordances learning.

2. Methods

As described in the Introduction, we trained our multilayer neural network to associate visual objects and actions, i.e., to form stable affordances. Each visual object was a rectangular bar defined by three features: orientation, colour, and shape. Remarkably, only one of these features at a time was relevant to choose the right action. To simulate the learning of affordances in different contexts, we trained our neural network in two subsequent phases or contexts: in the first context (natural), the network was exposed to a series of input patterns, including a visual object and

the corresponding desired action based on its orientation; afterwards, in the second context (artificial), the same network trained in the first context was exposed to a new series of input patterns in which the relevant feature co-varying with the desired action was alternatively the colour or the shape. In such a manner, we simulated the passage from a real-life context, in which a visual feature of the object was physically and strongly linked to the afforded action, to a laboratory-like context, in which the correspondence between object feature and action could be arbitrarily chosen.

The model and learning processes were implemented in MATLAB 7.10, version R2010a. The script was written with basic MATLAB commands, without using external libraries or toolbox. This software was running on a PC equipped with a Core i7-3770 3,4 GHz with 8 GB of DDR3 Ram. The computation was performed through the PC processor, without exploiting the GPU processors. With this system, model setup required about a minute, while each learning phase required from three to six hours, depending on the number of stimuli presented and epochs.

2.1. Inputs

Each pattern presented to the network (see **Figure 1** bottom, for an example) included two separate inputs: a visual input, consisting of a rectangular bar presented in one out of two 50x50 pixels matrices sensible respectively to the red and the green, and a corresponding action input, consisting of a 50x4 pixel pattern presented in a 50x40 pixels matrix. In particular, the visual input consisted of a red or green bar with an orientation in the range [1°, 90°]. The bar could be thin and long, from now on indicated as a bar with a 'thin' shape, with a width of 2 or 4 pixels and a height of 28 or 30 pixels, or thick and short, from now on indicated as a bar with a 'thick' shape, with a width of 10 or 12 pixels and a height of 20 or 22 pixels. The exact values for the width and height of each bar were randomly extracted between each pair of alternatives according to its defined shape. The action input consisted of a horizontal white line located at a variable height. During the first training context, the location of the white line varied within 90 possible positions depending on the orientation of the visual input. During the second training context, instead, the height of the bar varied with the colour or shape of the visual input. Perceptual noise was simulated by flipping the state of 2% of randomly selected units. The input stimuli were presented to the first layer of the network as they were, without any further preprocessing stage.

2.2. Neural network

We used a deep multilayer neural network [35], including bottom-up and top-down connections [36]. The ability of these networks to build a hierarchy of progressively more complex distributed representation and to generate sensory data in addition to

classifying them makes these data particularly attractive for neurocognitive modelling [36], [37].

The network included one visible layer and three hierarchically organized hidden layers, with each couple of layers connected through bottom-up (recognition) and top-down (generative) weights. Visual and action information was processed within two separate neural pathways within the first two layers and then combined in the third layer.

For the neural pathway encoding the visual information, we had two sets of red and green units arranged in two 50x50 matrices in the visible layer (for a total of 5000 units), enabling the detection of the presence and position of red and green stimuli. The first and second hidden layers included, respectively, 500 and 250 neural units. For the neural pathway that encodes the action information, we had an input matrix of 50x40 neural units in the visible layer and 250 and 125 neural units, respectively, in the first and second hidden layers. Finally, the third (common) hidden layer included 200 neural units (see **Figure 1**, Top).

This kind of neural network could be conveniently reduced to a stack of restricted Boltzmann machines (RBM) [38], in which a layer of feature detectors (hidden units) h_j received weighted input $x_j = \sum w_{ij} v_i$ from the previous layer. The activation of each unit in the feature detector was computed by passing the input through the logistic function $h_j = 1/(1+e-x_j)$.

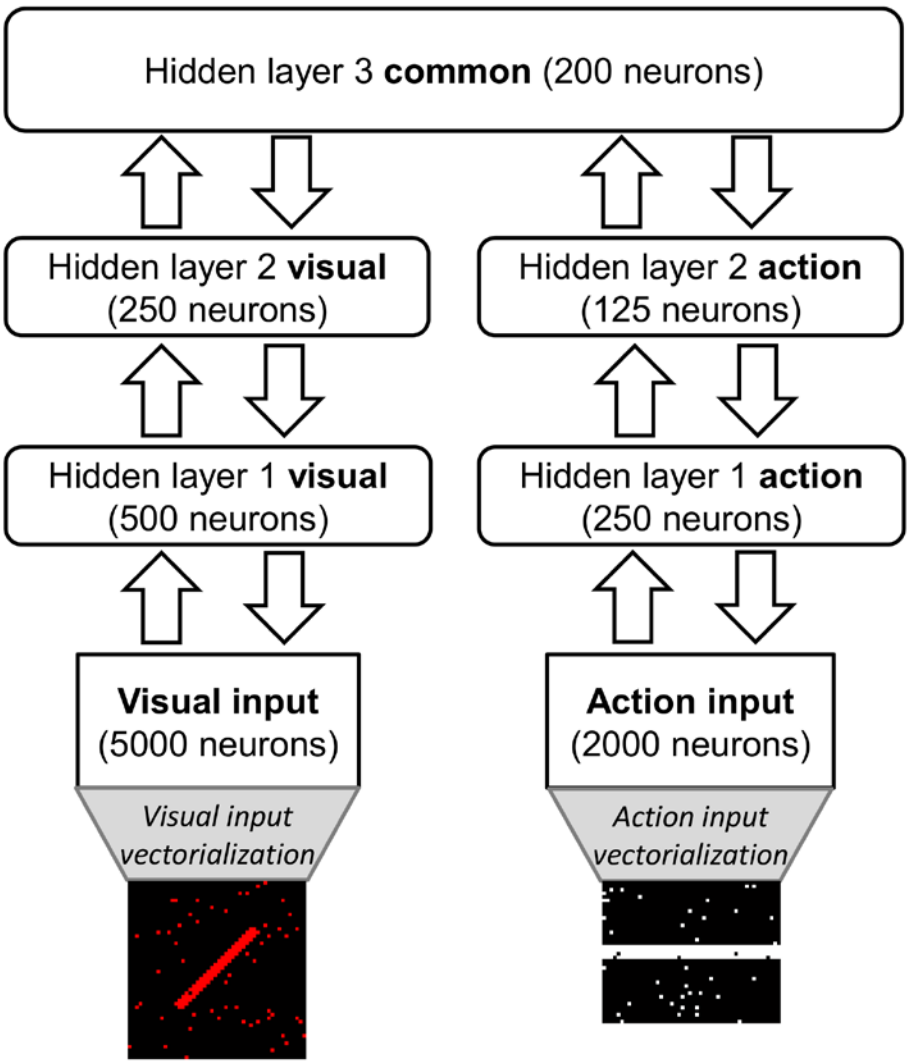

Figure 1. Top: The neural network architecture included a visible layer and three hierarchically organised hidden layers, with two separate neural pathways for encoding the visual and action information up to the second hidden layer. The third hidden layer received as input the combined output of the two pathways. The network layers were fully connected through both bottom-up (upward arrows) and top-down (downward arrows) weights. Bottom: An example of an input pattern including a thin red bar presented with an orientation of 45° and the corresponding action (i.e., a bar located halfway between the upper position, corresponding to an orientation of 1°, and the lower position, corresponding to an orientation of 90°).

2.3. Learning algorithm and validation procedure

The learning process was carried out through the unsupervised auto-associative

learning algorithm described in [36]. The pseudocode of the implied training algorithm is reported in Table 1. The connection weights were initially randomly set from a normal distribution (μ=0.0, σ=1) and then scaled by a factor of 0.1, and the neuronal biases were initially set at zero. The network was trained to generate the sensory data, i.e., to maximize the likelihood of reconstructing the input data, starting from a given state of the feature detectors and using the weights w_{ji} in a top-down direction. We first trained the two hidden layers in the visual neural pathways, then the two hidden layers in the action neural pathways, and finally the third layer. Each layer was trained for 200 learning epochs, and each learning epoch comprised the entire training set (see the next section).

For each hidden layer to train, given an input vector vi+, the activation of the feature detectors h_j^+ ("positive" phase) was first calculated. Starting from stochastically selected binary states of the feature detectors (using their state h_j^+ as a probability to turn them on), it then inferred an input vector v_i^- used in turn to reactivate the features detectors h^{j-} ("negative" phase). The weights w_{ij} were updated with a small learning fraction ε of the difference between the input-output correlations measured in the positive and negative phases: $\Delta w_{ji} = \varepsilon (v_i^+ h_j^+ - v_i^- h_j^-)$. The Δw_{ji} computed with this equation was then corrected considering the momentum of the previous gradient step η and the weight cost *c*. Neuron biases were also upgraded as well with ε. The learning parameters were as follows: $\varepsilon = 0.1$, $\eta = 0.5$, and $c = 0.0002$. All parameters were selected based on relevant previous papers, including this model [36].

The training dataset consisted of 10 input patterns for each combination of colour (red or green), orientation (from 1° to 90°) and shape (thin or thick), for a total of 3640 input patterns. Each input pattern included 7000 bits that specified the state of the corresponding visible units (that could assume either a 1.0 or 0.0 activation state), of which 5000 for the visual input and 2000 for the action input. To speed up the training process, we divided the training set into 10 batches of 360 input patterns, each of which included one pattern for each combination of the three visual features. After training, neural network validation was performed using partial inputs instead of complete input patterns in the testing phase. In practice, we presented to the neural network either the visual stimulus only or the action stimulus only for each training input, and assessed the capability of the network to complete the missing part of the inputs by exploiting its generative abilities. The network's responses to testing stimuli with missing input would be analysed and interpreted in terms of generalization of the learned visual-action associations.

Table 1. Pseudocode for the basic deep learning algorithm implied.

1. **Start**
2. **Initialize** the neural network: visual pathway (5000-500-250 neurons), action pathway (2000-250-125 neurons), and common hidden layer (500 neurons).
3. **Initialize** the weights of the network randomly from a normal distribution (μ=0.0, σ=1) and scale them by a factor of 0.1.
4. **Initialize** the bias of neurons = 0
5. **For each** hidden layer in the neural network to be trained:
6. **For each** training epoch:
7. **For each** input vector vi+:
8. **Computed** the activation of the feature detectors h_j^+ ("positive" phase) based on vi+
9. **Inferred** an input vector v_i^-
10. **Reactivate** the feature detectors h^{j-} ("negative" phase) based on v_i^-
11. **Compute** the difference in activation between the positive and negative phases.
12. **Update** weights and biases
13. **Save** the learned weights and biases of the trained layer.
14. **End**

3. Results

In this section, we described the behaviour of the neural network at the end of the training process in a normal condition in which the network experienced the same type of stimuli experienced during training and in testing conditions in which the stimuli were manipulated to verify the generalization capability of the network. Furthermore, to identify the effect of learning new associations, we analysed the behaviour of the network at the end of the first and second training phases. For each testing condition, we computed the average activation overall of 100 input patterns.

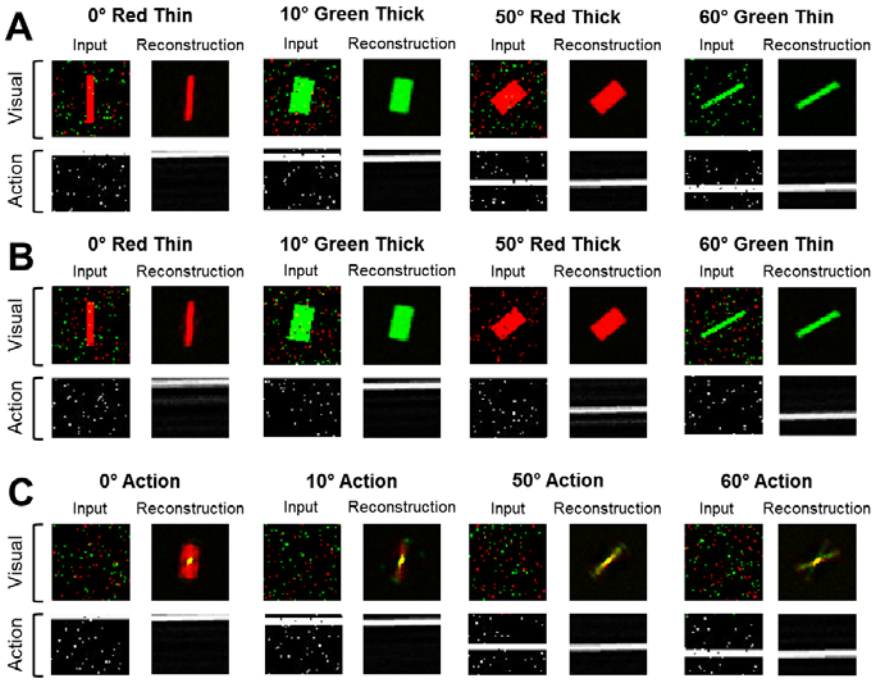

Figure 2. Example of input and self-generated data (left and right of each column, respectively) obtained in the normal condition in which the network received both the visual and action input (panel A) and in the test condition in which the network received only the visual input (panel B) or the action input (panel C).

3.1. Quality of data reconstruction and generalization

Analyzing the behaviour of the network at the end of the first training phase under normal conditions, we observed that it correctly reconstructed both visual and action inputs and filtered out noise (see **Figure 2A** for exemplific cases). The quality of reconstructions of both visual and action input was further demonstrated by measuring the average activity of units coding for correct and incorrect visual and action stimuli (**Figure 3**, left panel) during the presentation of 400 stimuli (including 100 stimuli for each of the four possible combinations of colour and shapes of the bar). When the action input was missing (**Figure 3**, central panel), for the action activity, we averaged the activation of the units coding for the expected position of the action based on the bar orientation; while when the visual input was missing (**Figure 3**, right panel), for the bar activity, we averaged the activation of the units coding for the expected positions of a small (width = 2, height = 20) red bar and a small green bar.

The analysis of the behaviour under test conditions, in which the network received only visual stimuli as input (**Figure 2B** and **Figure 3**, central panel) or only action

stimuli (**Figure 2C** and **Figure 3**, right panel), indicates that the network was able to generalize by partially regenerating also the missing stimuli. As exemplified in **Figure 2C**, when the visual stimulus was missing, and consequently the colour of the visual stimulus was unknown, the network tended to generate both a red and a green stimulus with the appropriate orientation.

In general, these results indicated that the network was able to learn the association between the relevant feature of the visual input, i.e., its orientation, and the appropriate action. Thus, the training process led to the formation of strong affordances automatically elicited by the orientation of the visual stimulus.

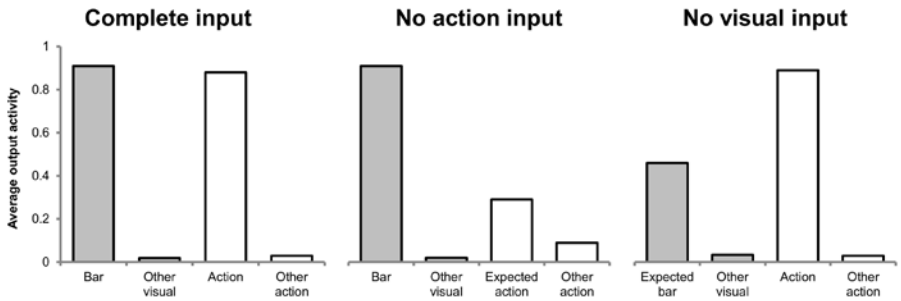

Figure 3. Average activity of the units coding for visual and action stimuli in the data generated by the network. Data are shown separately for the units encoding the position of the bar and the action, and for the remaining visual and action units. When an input (visual or action) was missing, we averaged the activation of the units coding for its expected position. Left panel: results in a normal condition in which the network experienced 400 input patterns, including both stimuli. Central panel: results obtained in a test condition in which the network experienced the same input patterns but in which the activation of all action units was set to 0.0. Right panel: results obtained in a test condition in which the network experienced the same items, but in which the activation of all visual units was set to 0.0.

3.2. Learning new associations based on colour and shape

As mentioned above, during the second training context, the network was further trained on stimuli in which the action correlated with the colour or with the shape of the visual input rather than with the visual input orientation. More specifically, in context 2A, the action varied according to the colour of the visual stimuli, with the red bars requiring a 1°-like action and the green bars requiring a 90°-like action, whereas, in context 2B, the action varied according to the shape of the visual stimuli, with thin bars requiring a 1°-like action and thick bars requiring a 90°-like action. The orientation of the bars was randomly extracted in the range [1°, 90°]. During these second training contexts, the network was initialized with the connection weights obtained at the end of the previous training context (i.e.,

the network previously trained in the first context was further trained in context 2A or context 2B). The new training context involved 10 epochs, including 3600 input patterns (i.e., 900 input patterns for each combination of colours and shapes). The 3600 input patterns presented during one epoch were divided into 10 batches of 360 units, including 90 combinations of colours and shapes.

To analyze the course of the new learning context, we tested the network after each training epoch on complete and incomplete input patterns (on input patterns that included visual and action inputs or only visual input only). In particular, we tested the network with input patterns affording different actions depending on the context, i.e., a red bar with an orientation of 90° associated with a 90°-action in context 1, and with a 1°-action in context 2A. By analysing how the network responded to complete input patterns, we observed that it correctly reconstructed both stimuli belonging to the first and second learning context after one epoch of context 2 training (data included in **Figure 7**).

More interesting, by analysing how the network reacted to incomplete stimuli that included visual but not action input, we observed that the network tended to regenerate complete stimuli that also included missing action (**Figure 4**). In this case, the network tended to generate actions that were consistent with the regularities experienced during the first context up to trials 4 and 9, respectively (in the case of context 2A and 2B, respectively) and with the regularities experienced during the second training context afterwards (**Figure 4** top and bottom panel, respectively).

In general, these data indicated that the network can extract and incorporate the regularities that characterized the second training context while preserving those that characterized the first training process. Moreover, they indicated that in the case of incomplete input patterns (in which the appropriate context cannot be inferred from the current combination of visual and action inputs), the network tended to respond preferentially based on the contexts characterizing recently experienced inputs. Finally, learning of new regularities required more training epochs in the case of context 2B (shape-based learning) concerning context 2A (colour-based learning), indicating that the acquisition of the new regularities was easier when such regularities concerned the colour rather than the shape.

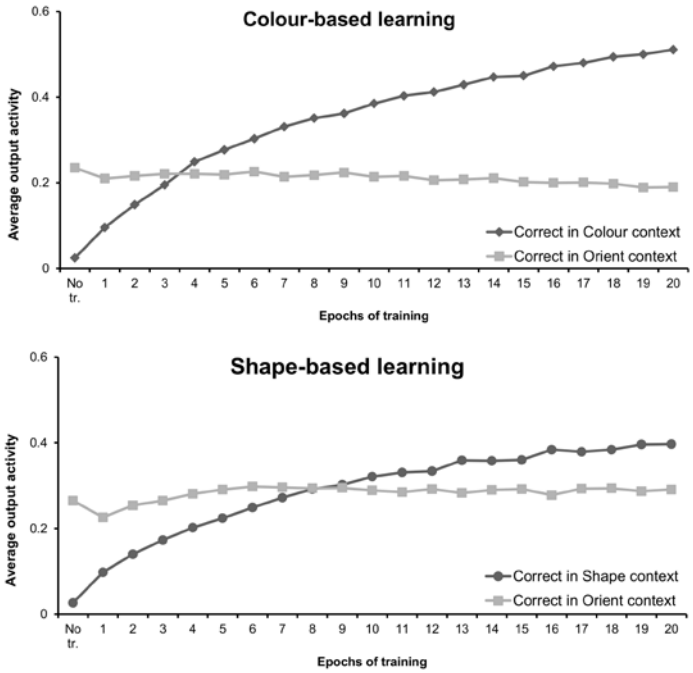

Figure 4. Average action activity of the units coding for the action appropriate in the first ('Correct in Orient context') and second ('Correct in Colour context' or 'Correct in Shape context') context in response to incomplete input patterns after each epoch of training in the second context. Top panel: data obtained in the case of context 2A in which the new action co-varied with the colour of the visual input. Bottom panel: data obtained in the case of context 2B in which the new action co-varied with the shape of the visual input.

3.3. Resolving conflicts between multiple afforded actions

As reported in the previous sub-section, the neural network learned to associate the colour or the shape of the visual input with the action. However, it also continued to reconstruct the action associated with orientation even after 20 learning epochs in the second context (as shown in **Figure 4**). Interestingly, the network showed an ability to spontaneously converge on the most plausible action in response to stimuli that offered different alternative actions when it was allowed to process each stimulus for multiple time steps. To enable the neural network to process information over multiple time steps, we allowed it to operate for the first computational step by activating the input units based on the externally provided information and for the further computational steps based on the self-generated state by the network during the previous time step. As in the previous sub-section, we tested the neural network during the colour-based and shape-based learning phases with input patterns affording two alternative actions co-varying with the

regularities experienced respectively during the first and during the second context learning phases.

The results of these simulations are reported in **Figure 5**, in which the activation of the units that encode the two actions is plotted against the number of learning epochs and over three successive computational steps for both the colour-based learning phase (panel A) and the shape-based learning phase (panel B). As shown, in the first learning epochs, the neural network regenerations of the two actions were differentiated over time, the action appropriate in context 1 was increasingly activated while that appropriate in context 2 was decreasingly activated from the first to the third computational step. As the training progressed, this pattern reversed, with increasingly more activation of the action appropriate in context 2 over time steps, while the action appropriate in context 1 had no increase in activity. Then, in the last training epochs, the action appropriate in context 1 even lost activation over time steps, especially in the colour-based learning condition (panel A).

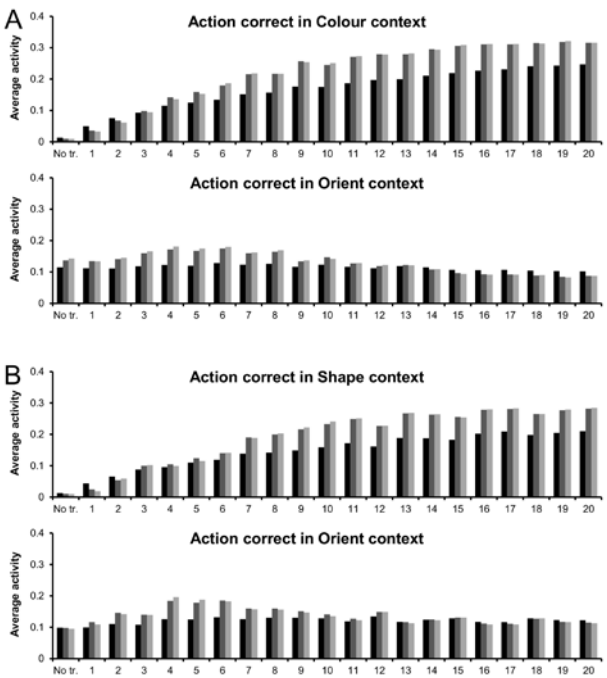

Figure 5. Average activation of the actions correlated with the first and second learning contexts, during three successive computational steps (shown in black, grey and light grey) after each of the 20 epochs of context 2 learning. These data were obtained by using incomplete input patterns, including only visual input. Panel A contrasts the results obtained in the colour-based and orientation-based learning

phase (context 2A and context 1), while panel B contrasts the results obtained in the shape-based and orientation-based learning phase (context 2B and context 1). Bottom panel: data obtained in the case of context 2B in which the new action co-varied with the shape of the visual input.

3.4. Availability of colour versus shape representations

As shown in Section 3.2, learning new associations between colours and actions was faster and better than learning new associations between shape and actions. One possible explanation of this effect could be that colour information was more readily available in the neural network after the first context learning with respect to shape information.

To verify this hypothesis, we estimated to what extent stimuli of different orientations, colours, and shapes were differentiated in the network representations at the end of the first training phase. This was realized by training a linear neural network including two units that received and projected connections from and to the third hidden layer of the network. This additional layer was trained for 10 epochs with 3640 input patterns, including an equal number of visual inputs for each possible combination of two orientations (1° or 90°), two colours (red or green), and two shapes (thin or thick) of the bar. Action inputs were not provided. The connection weights of the linear network were randomly initialized in the range [0.0, 0.1], the neuronal biases of the two neurons were set to zero, and the learning rate was set to $\varepsilon = 0.001$. Indeed, as a result of auto-associative training, activation of these two units can be used to roughly characterize in a low-dimensional space the overall distribution of the representation of stimuli in network modelling [36], [37].

As can be seen in **Figure 6**, the state of these two additional neurons for visual inputs with different colours, orientations, and shapes, visual inputs varying with respect to colour were much more separated in the representational space than visual inputs that varied with respect to shape (**Figure 6** left and right panel). These data were confirmed by the geometric separability index analysis [39], that is, by the analysis of the proportion of stimuli that had as their nearest neighbour (in the two-dimensional space of the linear network two-dimensional space) a stimulus of their same category. This index, which varied between 0.5 and 1.0, corresponding, respectively, to randomly overlapping and fully separated distributions, was 0.97 in the case of visual input with different colours and 0.78 in the case of visual input with different shapes (see **Figure 6**). The representations of different coloured visual inputs were also more separated from those of different orientated visual inputs (0.87).

These results thus confirmed that the faster acquisition of colour/action

relationship with respect to shape/action relationship could be explained by the different levels of availability of categorical colour information with respect to categorical shape information in the network representations and that, in principle, the learning of colour-based actions should be easier than the learning of orientation-based actions.

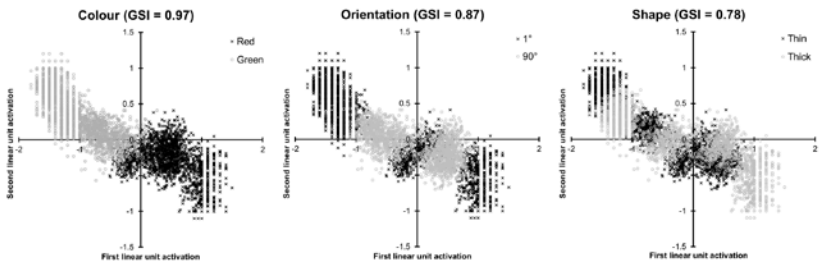

Figure 6. Geometric separability index and state of the two neurons of the linear neural network for visual inputs with different colours (left panel), orientation (middle panel), and shapes (right panel).

3.5. Learning context-dependent associations

In this last experimental section, we report the data obtained in a series of experiments in which we subjected the neural network to a training process during which the learning context periodically switched from context 1 (orienting-based learning) to context 2 (colour-based or shape-based learning). In this manner, we evaluated whether the neural network was able to select the action correct to the actual context and if the alternation between the two contexts had a cost or a benefit for such context-based action selection. As we did for the previous analysis, we tested the network reconstructions of the actions after each learning epoch by computing the average output activation of the units coding for the action appropriate to the current or to the alternative context with complete and incomplete input patterns (i.e., with stimuli including both the visual and action inputs or only the visual input).

Figure 7 reports the output activity relative to the appropriate action in the orientation-based training (grey line) and the colour-based (A panel) or the shape-based (B panel) training (black line) tested during the orientation-based training phase (light grey area) or the colour- and shape-based training phase (dark grey area). As shown, when the input patterns included the action input, the network regenerated only the experienced input action and did not activate at all the action that would be appropriate for the other context.

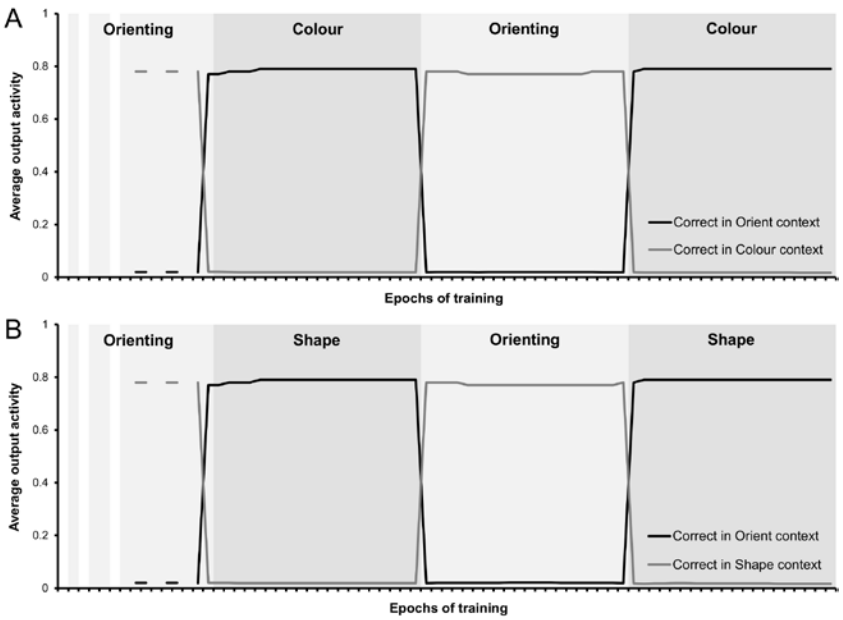

Figure 7. Actions generated by the neural networks during four consecutive learning phases in which the context switched from orientation-based to colour-based (panel A) or from orientation-based to shape-based (panel B) and vice versa, and in which the network was tested with complete input patterns.

Figure 8 instead reports the same analysis performed in a case in which the network was tested with incomplete input patterns, in which the action input was missing. As already pointed out in Section 3.2, under a condition in which the appropriate context could not be inferred from the experienced stimulus, the network tended to activate both actions (i.e., the action that was appropriate for the orientation-based context and the action that was appropriate for the colour-based or the shape-based context). Moreover, the network was also able to suppress or increase the strength of the associations through training repetitions based on the actual training phase context, indicating that it needed some training epochs to switch its preference from one learning context to the other.

Finally, the comparison of the variation in network behaviour during the first and second switch from one context to another (**Figure 8**) indicated that the training time necessary to switch this preference becomes shorter during successive context alternations. Moreover, the difference in activation between the two reconstructed actions increased in the second switch from context 1 to both context 2A and 2B, with a better reconstruction of the appropriate action and poorer reconstruction of the alternative one.

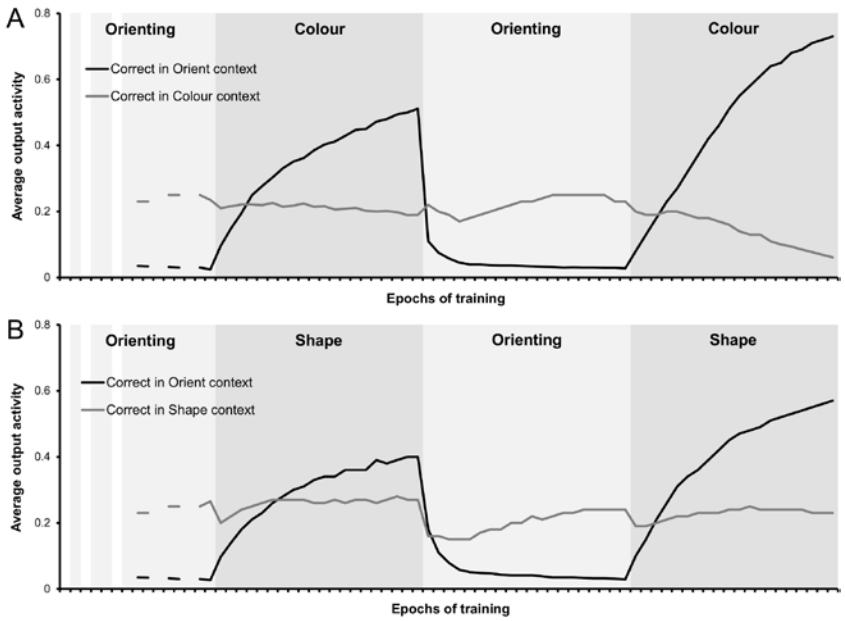

Figure 8. Actions generated by the neural networks during four consecutive learning phases in which the context switched from orienting-based to colour-based (panel A) or from orienting-based to shape-based (panel B) and vice versa, and in which the network was tested with incomplete input patterns, without the action input.

4. Discussion, Opportunities, and Open Issues

The present study allowed us to investigate the processes underlying the formation of affordances, i.e., of novel associations between perceptual properties (orientation, shape, and colour) and action. Using artificial neural networks, we were able to manipulate and analyse all possible variables and avoid biases induced by the previous knowledge of participants that inevitably affect the data collected with human subjects.

Our study had two objectives: the first was to test with neural networks whether affordances are automatically activated or contextual dependent, and the second was to verify, once orientation-action associations are established, whether colour-action associations are easier to learn compared to shape-action associations. We will discuss below how the two objectives were addressed and reached.

As to the first objective, our results have clear implications for the literature on affordances and the debate concerning their automaticity. They indeed suggest that the views according to which affordances are automatic and those according to which they are completely context- and task-dependent should be reconciled [11], [40]. As already described in the Results section, when the neural network was submitted to

the new learning phase, i.e., once it was trained to associate an action with colour or shape, the test of network reconstructions revealed that the old associations were not lost (see **Figure 4**). The average output activity of the units coding the old action remained substantially stable. The process occurred over time (see **Figure 5**). The pattern of these results indicates that context plays a role in affordance activation. At the same time, however, the results do not support the view that the context/task selects only the relevant affordances: the old action remained activated over time, even if its influence progressively decreased. Overall, our results favor the view that automaticity and context/task-based selection are not incompatible. All affordances, related to both old and new actions, are automatically activated. However, the context/task modulates the strength of such automatic activation, i.e., actions coherent with the ongoing context/task are activated stronger than actions incoherent or partially coherent with the ongoing context/task. Hence, the context/task seems to act as a sort of late filter that selects the association that is more relevant to the current action/goal. This is consistent with a view according to which first all object affordances are activated and compete with each other, and then the affordances relevant to the current context are recruited. So, our simulations differ from many previously proposed neural network models of action selection, in which the context causes the suppression of irrelevant visual information to block the instantiation of inappropriate action [30] or allows filtering affordances after their instantiation through a competitive mechanism [24], [41]. In our neural network simulations, the context/task modulates both the initial strength of the afforded action representations ('early stage') and their subsequent selection based on their actual relevance ('late stage').

As to the second objective, our results reveal that not all contexts and tasks have the same impact. Consistent with our predictions, we found that learning colour-based associations was easier than learning shape-based associations. This result is clearly in line with the inspiration provided by the experimental evidence [27], although in the present study we investigated how learning new colour-action associations or shape-colour associations changed after having previously learned orientation-action associations. Instead, the experiments simply used tasks in which participants had to perform decisions on the object's colour vs. orientation/shape. A further major difference between the present study and other experimental studies was that, with neural networks, no pre-existing bias toward the activation of a specific perceptual feature was present. Specifically, our results show that the pattern of activation of the units coding the old and new actions diverges for shape-based and color-based associations since the new action activation increases and the old action activation decreases were more pronounced for colour-based associations. Furthermore, our

analyses allowed us to detect the mechanisms underlying the learning ability of the different associations. While colour categories are separate, the distinction between shape categories is less marked.

Our study had no direct technological applications, nor it was applied to robot models or simulated effectors. However, we showed that it would be possible to have a system that learns on the go new affordances as associations between a certain visual stimulus, a certain action, in a certain environmental situation or task. We showed how this plastic system could be trained both offline and online (see Section 3.5). Future studies could apply this neural network to robotic agents such as they would not only learn object-action pairs but also to conveniently select the appropriate affordance in a given context. While training large deep neural networks could be costly, maybe mixed solutions using both an offline learner for the affordances (object-action pairs) and then an online learner, collecting information about the context and selecting the best affordance in such a context.

We used as a neural network model the one proposed by Hinton [36]. However, other models and learning rules could be tested with the same data, such as recurrent neural networks or deep convolutional networks [35], [42]. Future works could test and compare such alternative models to advance our understanding of human cognitive processes. Also, a more compelling technical implementation could be tested, relying on GPU processors or parallel processing. This could reduce the learning time required for the model and then make a robotic or other technological application of our work based on online computing more viable.

Unlike most robotic or artificial neural network affordance learning (see [34]), we did not use a prior knowledge base or human-coded affordances. Instead, we trained our neural network directly with pairs of visual objects and desired actions, in accordance with the Gibson proposal of direct affordance perception independent of categorical or verbal knowledge [1]. This method allowed the development of robust and context-dependent affordances in terms of visual-action associations. As our trained neural network replicated some key characteristics of the human affordance system, this work could be conveniently applied to different fields in which an artificial platform could be requested to have such characteristics. For example, a revised version of our model could be applied to behaviour-based robotics, which relies on direct perception [43], or to human-robot interactions and educational contexts, in which shared affordances between humans and artificial systems would allow a more direct and fruitful interaction [44]. Lastly, the present model could also be tested in a more complex scenario in which learned affordances would be selected based on a specific context that changes over time [30], [45].

Overall, these results suggest that collecting experimental data on human subjects

concerning the dynamics of the process during which new associations are formed could significantly extend our understanding of these phenomena. More generally, the combined use of artificial neural network models, which can be trained from scratch, easily manipulated, and analysed, with experimental data, as in [46], can represent a powerful method to better understand how affordances are formed and activated.

Conflicts of Interest

All authors declare that they have no conflicts of interest.

References

[1] J. J. Gibson, *The Ecological Approach to Visual Perception*, vol. 40, no. 1. Boston: Houghton Mifflin, 1979. doi: 10.2307/989638.
[2] S. Thill, D. Caligiore, A. M. Borghi, T. Ziemke, and G. Baldassarre, "Theories and computational models of affordance and mirror systems: An integrative review," *Neurosci Biobehav Rev*, vol. 37, no. 3, pp. 491–521, Mar. 2013, doi: 10.1016/j.neubiorev.2013.01.012.
[3] L. W. Barsalou, "Grounded cognition.," *Annu Rev Psychol*, vol. 59, pp. 617–45, Jan. 2008, doi: 10.1146/annurev.psych.59.103006.093639.
[4] A. M. Borghi and D. Pecher, "Introduction to the special topic embodied and grounded cognition.," *Front Psychol*, vol. 2, p. 187, Jan. 2011, doi: 10.3389/fpsyg.2011.00187.
[5] J. I. Davis and A. B. Markman, "Embodied cognition as a practical paradigm: introduction to the topic, the future of embodied cognition.," *Top Cogn Sci*, vol. 4, no. 4, pp. 685–91, Oct. 2012, doi: 10.1111/j.1756-8765.2012.01227.x.
[6] G. Dove, "Beyond the body?The future of embodied cognition." Frontiers in Psychology, 2014.
[7] H. E. Matheson and L. W. Barsalou, "Embodiment and Grounding in Cognitive Neuroscience," *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, pp. 1–27, Mar. 2018, doi: 10.1002/9781119170174.EPCN310.
[8] M. Tucker and R. Ellis, "On the relations between seen objects and components of potential actions.," *J Exp Psychol Hum Percept Perform*, vol. 24, no. 3, pp. 830–846, Jun. 1998, doi: 10.1037/0096-1523.24.3.830.
[9] M. Tucker and R. Ellis, "The potentiation of grasp types during visual object categorization.," *Vis cogn*, vol. 8, pp. 769–800, 2001, doi: 10.1080/13506280042000144.
[10] M. Tucker and R. Ellis, "Action priming by briefly presented objects.," *Acta Psychol (Amst)*, vol. 116, no. 2, pp. 185–203, Jun. 2004, doi: 10.1016/j.actpsy.2004.01.004.
[11] M. van Elk, H. van Schie, and H. Bekkering, "Action semantics: A unifying conceptual framework for the selective use of multimodal and modality-specific object knowledge," *Phys Life Rev*, vol. 11, no. 2, pp. 220–50, Jun. 2013, doi: 10.1016/j.plrev.2013.11.005.
[12] M. Costantini, E. Ambrosini, G. Tieri, C. Sinigaglia, and G. Committeri, "Where does an object trigger an action? An investigation about affordances in space.," *Exp Brain Res*, vol. 207, no. 1–2, pp. 95–103, Nov. 2010, doi: 10.1007/s00221-010-2435-8.

[13] M. Costantini, E. Ambrosini, C. Scorolli, and A. M. Borghi, “When objects are close to me: affordances in the peripersonal space.,” *Psychon Bull Rev*, vol. 18, no. 2, pp. 302–8, Apr. 2011, doi: 10.3758/s13423-011-0054-4.
[14] E. Y. Yoon, G. W. Humphreys, and M. J. Riddoch, “The Paired-Object Affordance Effect.,” *J Exp Psychol Hum Percept Perform*, vol. 36, no. 4, pp. 812–824, Jul. 2010, Accessed: Jan. 22, 2015. [Online]. Available: http://eric.ed.gov/?id=EJ894200
[15] A. M. Borghi, A. Flumini, N. Natraj, and L. A. Wheaton, “One hand, two objects: emergence of affordance in contexts.,” *Brain Cogn*, vol. 80, no. 1, pp. 64–73, Oct. 2012, doi: 10.1016/j.bandc.2012.04.007.
[16] S. Kalénine, A. D. Shapiro, A. Flumini, A. M. Borghi, and L. J. Buxbaum, “Visual context modulates potentiation of grasp types during semantic object categorization.,” *Psychon Bull Rev*, vol. 21, no. 3, pp. 645–51, Jun. 2014, doi: 10.3758/s13423-013-0536-7.
[17] C. Lee, E. Middleton, D. Mirman, S. Kalénine, and L. J. Buxbaum, “Incidental and context-responsive activation of structure- and function-based action features during object identification.,” *J Exp Psychol Hum Percept Perform*, vol. 39, no. 1, pp. 257–70, Feb. 2013, doi: 10.1037/a0027533.
[18] F. Ferri, G. C. Campione, R. Dalla Volta, C. Gianelli, and M. Gentilucci, “Social requests and social affordances: how they affect the kinematics of motor sequences during interactions between conspecifics.,” *PLoS One*, vol. 6, no. 1, p. e15855, Jan. 2011, doi: 10.1371/journal.pone.0015855.
[19] L. Sartori, C. Becchio, M. Bulgheroni, and U. Castiello, “Modulation of the action control system by social intention: unexpected social requests override preplanned action.,” *J Exp Psychol Hum Percept Perform*, vol. 35, no. 5, pp. 1490–500, Oct. 2009, doi: 10.1037/a0015777.
[20] C. Scorolli, M. Miatton, L. A. Wheaton, and A. M. Borghi, “I give you a cup, I get a cup: a kinematic study on social intention.,” *Neuropsychologia*, vol. 57, pp. 196–204, May 2014, doi: 10.1016/j.neuropsychologia.2014.03.006.
[21] A. M. Borghi, “Affordances, context and sociality,” *Synthese*, vol. 199, no. 5–6, pp. 12485–12515, Dec. 2021, doi: 10.1007/s11229-018-02044-1.
[22] M. Mustile, F. Giocondo, D. Caligiore, A. M. Borghi, and D. Kourtis, “Motor Inhibition to Dangerous Objects: Electrophysiological Evidence for Task-dependent Aversive Affordances,” *J Cogn Neurosci*, vol. 33, no. 5, pp. 826–839, Apr. 2021, doi: 10.1162/jocn_a_01690.
[23] M. L. Kellenbach, M. Brett, and K. Patterson, “Actions speak louder than functions: the importance of manipulability and action in tool representation.,” *J Cogn Neurosci*, vol. 15, no. 1, pp. 30–46, Jan. 2003, doi: 10.1162/089892903321107800.
[24] P. Cisek, “Cortical mechanisms of action selection: the affordance competition hypothesis.,” *Philos Trans R Soc Lond B Biol Sci*, vol. 362, no. 1485, pp. 1585–99, Sep. 2007, doi: 10.1098/rstb.2007.2054.
[25] G. Pezzulo and P. Cisek, “Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition,” *Trends Cogn Sci*, vol. 20, no. 6, pp. 414–424, Jun. 2016, doi: 10.1016/j.tics.2016.03.013.
[26] S. P. Tipper, M. a Paul, and A. E. Hayes, “Vision-for-action: the effects of object property discrimination and action state on affordance compatibility effects.,” *Psychon Bull Rev*, vol. 13, no. 3, pp. 493–8, Jun. 2006, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17048736.

[27] A. Pellicano, C. Iani, A. M. Borghi, S. Rubichi, and R. Nicoletti, “Simon-like and functional affordance effects with tools: The effects of object perceptual discrimination and object action state,” *Q J Exp Psychol*, vol. 63, no. 11, pp. 2190–201, 2010, doi: 10.1080/17470218.2010.486903.
[28] G. M. Anderson, D. Heinke, and G. W. Humphreys, “Featural guidance in conjunction search: the contrast between orientation and color.,” *J Exp Psychol Hum Percept Perform*, vol. 36, no. 5, pp. 1108–27, Oct. 2010, doi: 10.1037/a0017179.
[29] G. Morlino, C. Gianelli, A. M. Borghi, and S. Nolfi, “Learning to Manipulate and Categorize in Human and Artificial Agents.,” *Cogn Sci*, Jul. 2014, doi: 10.1111/cogs.12130.
[30] L. Simione and S. Nolfi, “The Role of Selective Attention and Action Selection in the Development of Multiple Action Capabilities.,” *Conn Sci*, vol. 26, no. 4, pp. 389–402, 2014, doi: 10.1080/09540091.2014.942597.
[31] L. Simione and S. Nolfi, “Selection-for-action emerges in neural networks trained to learn spatial associations between stimuli and actions,” *Cogn Process*, vol. 16, pp. 393–397, 2015, doi: 10.1007/s10339-015-0679-8.
[32] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic, “Affordance detection for task-specific grasping using deep learning,” in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, Nov. 2017, pp. 91–98. doi: 10.1109/HUMANOIDS.2017.8239542.
[33] T.-T. Do, A. Nguyen, and I. Reid, “AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1–5. doi: 10.1109/ICRA.2018.8460902.
[34] N. Yamanobe *et al.*, “A Brief Review of Affordance in Robotic Manipulation Research,” *Journal of the Robotics Society of Japan*, vol. 36, no. 5, pp. 327–337, 2018, doi: 10.7210/JRSJ.36.327.
[35] C. C. Aggarwal, “Neural Networks and Deep Learning,” *Neural Networks and Deep Learning*, 2018, doi: 10.1007/978-3-319-94463-0.
[36] G. E. Hinton, “Learning multiple layers of representation.,” *Trends Cogn Sci*, vol. 11, no. 10, pp. 428–34, Oct. 2007, doi: 10.1016/j.tics.2007.09.004.
[37] M. Zorzi, A. Testolin, and I. P. Stoianov, “Modeling language and cognition with deep unsupervised learning: a tutorial overview.,” *Front Psychol*, vol. 4, no. August, p. 515, Jan. 2013, doi: 10.3389/fpsyg.2013.00515.
[38] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science (1979)*, vol. 313, no. July, pp. 504–507, 2006, Accessed: Oct. 15, 2013. [Online]. Available: http://www.sciencemag.org/content/313/5786/504.short.
[39] C. Thornton, “Separability is a learner’s best friend,” in *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, J. Bullinaria, D. Glasspool, and G. Houghton, Eds. London: Springer Verlag, 1997, pp. 40–47. doi: 10.1007/978-1-4471-1546-5_4.
[40] F. Binkofski and L. J. Buxbaum, “Two action systems in the human brain.,” *Brain Lang*, vol. 127, no. 2, pp. 222–9, Nov. 2013, doi: 10.1016/j.bandl.2012.07.007.
[41] P. Cisek and J. F. Kalaska, “Neural mechanisms for interacting with a world full of action choices.,” *Annu Rev Neurosci*, vol. 33, no. March, pp. 269–98, Jan. 2010, doi: 10.1146/annurev.neuro.051508.135409.
[42] G. H. de Rosa and J. P. Papa, “Soft-Tempering Deep Belief Networks Parameters Through Genetic Programming,” *Journal of Artificial Intelligence and Systems*, vol. 1, no. 1, pp. 43–59, 2019, doi: 10.33969/ais.2019.11003.

[43] R. Arkin, *Behavior-based robotics*. MIT Press, 1998. Accessed: Jul. 12, 2022. [Online]. Available: https://sci-hub.do/https://books.google.com/books?hl=it&lr=&id=mRWT6alZt9oC&oi=fnd&pg=PR11&dq=Arkin+R.+Behavior-based+Robotics.+MIT+Press.+1998&ots=46ZuhfVblx&sig=Dn0hmnMMLwiRYx5yqzzR2TUuOGA.

[44] T. Shu, M. S. Ryoo, and S.-C. Zhu, "Learning Social Affordance for Human-Robot Interaction," in *InternationalJoint Conference on Artificial Intelligence.*, 2016.

[45] L. Simione and S. Nolfi, "The emergence of selective attention through probabilistic associations between stimuli and actions," *PLoS One*, vol. 11, no. 11, p. e0166174, 2016, doi: 10.1371/journal.pone.0166174.

[46] A. Saxena, A. Khanna, and D. Gupta, "Emotion Recognition and Detection Methods: A Comprehensive Survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020, doi: 10.33969/AIS.2020.21005.