Specialization or Generalization: Investigating NeuroEvolutionary Choices via Virtual fMRI

Kevin Godin-Dubois[1,2] and Sylvain Cussat-Blanc[2,3] and Yves Duthen[2]

[1] Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[2] University of Toulouse, IRIT - CNRS UMR 5505, Toulouse, France
[3] Artificial and Natural Intelligence Toulouse Institute
k.j.m.godin-dubois@vu.nl

Abstract

Artificial Neural Networks have been crowned with tremendous successes in recent years, with ever wider and more complex ranges of applications. However, they, too often, result from a costly human design process relying as much on expertise as on trial and error. While the field of NeuroEvolution provides a complementary view point through emergent, self-designing ANNs, the "black-box" properties of the resulting networks is further magnified. Still, by once more taking inspiration from biology, we may extract meaningful information from ANNs by using similar approaches as those used for biological brains.

In this work, we study the emergence and functional allocation of neurons in a light communication task. By having a robot transmit visual information, through vocal channels, we enrich the existing literature with new types of stimuli, namely those related to role (emitter/receiver). Through Virtual functional Magnetic Resonance Imaging (VfMRI), we observe that evolution only favored specific kind of input-processing modules. Combined with a strong presence of jack-of-all-trades modules, this demonstrates the balancing act between specialization and generalization in Artificial Neural Networks with emergent topologies.

Introduction

One fascinating application of Artificial Neural Networks (ANNs) is as bio-mimetic engines that can reproduce critical characteristics of the animal brain, thereby paving the way towards its understanding (Treccani, 2020). However, while current ANNs can reach human performance on multiple abstract tasks such as playing Atari games (Mnih et al., 2015) or knowledge restitution (Bubeck et al., 2023), we are still far off Artificial General Intelligence (AGI). Some authors caution against assigning cognitive capabilities to Machine Learning agents, stating that "task-specific performance can [not] be treated as manifestation of General Intelligence" (Kadam and Vaidya, 2021). In parallel, it is argued in (Zador, 2019) that biological learning is not the result of clever algorithms but is likely, instead, to depend on a genomic bottleneck responsible for a brain's rapid learning capacities. The author states that "AI is far from achieving the intelligence of a dog or a mouse, or even of a spider [...]" at least in terms of general intelligence.

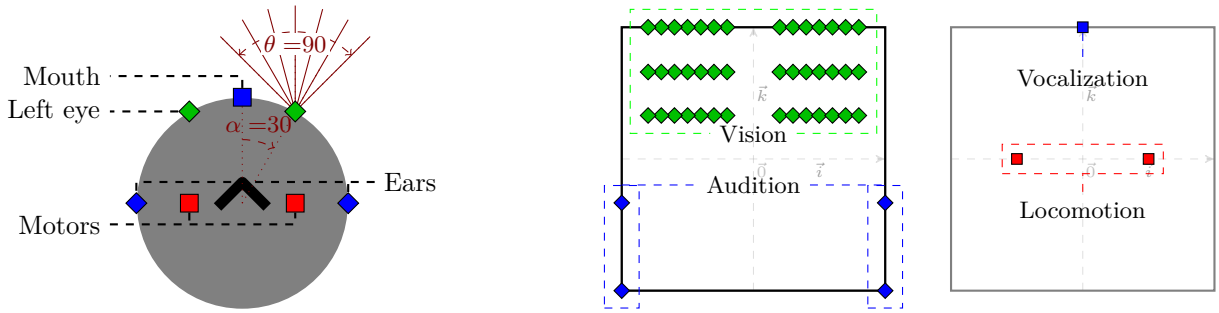
The research field of NeuroEvolution provides a way around this problem by taking inspiration not only from the biological brain but also from biological evolution (Stanley et al., 2019). Indeed, we argue that reaching the intelligence level of e.g. a spider requires grounded neural cognition to study, for instance, the evolutionary dynamics of vision (Olson et al., 2016) or communication (Kadish et al., 2019). The latter case could even lead to "natural" (non-)verbal communication with stepping stones including mimicking the bee's waggle dance (Campos and Froese, 2019) or primordial social dynamics (Ito et al., 2013).

The drawback of this approach, however, is that, by relinquishing control over an ANN's architecture, we further increase its inexplicability. While there has been increasing work done on relating Evolutionary Computation (EC) and Explainable AI (XAI), there is much that can be done for NeuroEvolution (Bacardit et al., 2022). Current investigations have covered different areas such as the evolution of explainable Hebbian learning rules (Mettler et al., 2021; Yaman et al., 2021) or interpretability through self-attention (Tang et al., 2020). However, neurons, their connectivity patterns and the role they occupy in a network have received far less attention, even though topology has been shown to be an important factor (Gaier and Ha, 2019).

In this work, we set out to study emergent neural structures in a light communication task where cooperation is essential. By placing two robots in disjointed arenas and having one communicating visual information to the other, we investigate whether individual neurons would tend to occupy specific functional niches. Formally the hypotheses are that the neurons will be separated by:

H1 function (visual versus auditory processing)

H2 role (differential activation based on situation)



(a) Body with inputs (diamonds) and outputs (squares).

(b) Neural substrate for inputs (left) and outputs (right).

Figure 1: Geometrical relationships between the robot's inputs, outputs and neurons. Left, the robot has both vision (green) and audition (blue). Eyes are placed at the front with overlapping fields and ears are on either side. Motors are placed closer to the center and the mouth (for vocalization) is at the front. Right, the physical positions are replicated in the neural substrate w.r.t. bilateral symmetry and relative placement.

The paper is organized as follows: Model presents the robots and ANNs, while Virtual fMRI introduces the methodology for extracting functional partitions. We then describe the Experimental protocol and some Population-level results, before detailing the actual Functional partitioning induced by this experiment.

Model

The virtual robots used in this experiment are based upon Godin-Dubois et al. (2023), without the morphological components, and are controlled by an ANN derived, through ES-HyperNEAT (Risi and Stanley, 2012), from a three-dimensional substrate[1].

A brief working knowledge of the platform is provided here. As illustrated in Figure 1a, these robots are solely composed of a circular body endowed with low-level perceptions and actions. Audition is implemented, in this work, via two channels (noise, resulting from motion; and explicit vocalization) with attenuation so that echolocation can emerge. Vision is performed through 7 ray-casts per eye, fanning at $\pi/2$, which form two rudimentary retinas. These parameters give the robot a moderate-grain forward facing visual field which is used to populate the corresponding neural layer with the RGB components of the first collided object. Actions are similarly elementary with both motors allowing for back and forward motion while vocalization is controlled through $volume(o) = max(0, o)$ where $o \in [-1, 1]$ is the corresponding neuron's output.

Emerging topologies

All of a creature's neurons are geometrically positioned in a 3D substrate, bounded in $[-1, 1]^3$, with two regions having hard-coded characteristics: the input and output planes placed at $y = -1$ and $y = 1$, respectively (Figure 1b). On the former, one neuron is allocated to each frequency (noise/voice) and ear while maintaining bilateral information. The retina is similarly modeled by a larger collection of neurons, each encoding the specific red, green or blue component of a particular ray for a particular eye. The angular position of the ray's endpoint is translated into the x coordinate of the neuron, while the color component gives the elevation z. The output plane comprises the three effective neurons: two for the motors and one for vocalization.

The rest of the neural controller is obtained through the ES-HyperNEAT algorithms. The core component of this methodology is a Composite Pattern Producing Network (Stanley (2007), CPPN) which is an $\mathbb{R}^n \to \mathbb{R}^m$ mathematical function composed of numerous elementary processing units $(\sin(x), e^x, |x|, ...)$. This function is then evolved in the same manner as the neural networks from NEAT (Stanley and Miikkulainen, 2002). Because direct encoding is not scalable to higher dimensional spaces, this CPPN is used indirectly, to describe connectivity patterns in a substrate.

As described in HyperNEAT (Stanley et al., 2009), the use of such an encoding leverages the CPPNs' capacity to capture repetition, symmetries and repetition with variations. However, one drawback of HyperNEAT is that it still requires that the experimenter manually places every hidden neuron, preventing any evolutionary adaptation. By opposition, the Evolvable Substrate extension (Risi and Stanley (2012), ES-HyperNEAT), takes advantage of the CPPN's connectivity pattern to automatically instantiate hidden neurons at locations of "high informational density".

This way, a single genomic component of unbounded complexity is responsible for the indirect encoding of the whole of the creature's brain: from the placement and density of hidden neurons to the topology and eventual emergence of neural structures.

[1]Source code: `https://github.com/kgd-al/Splinoids`

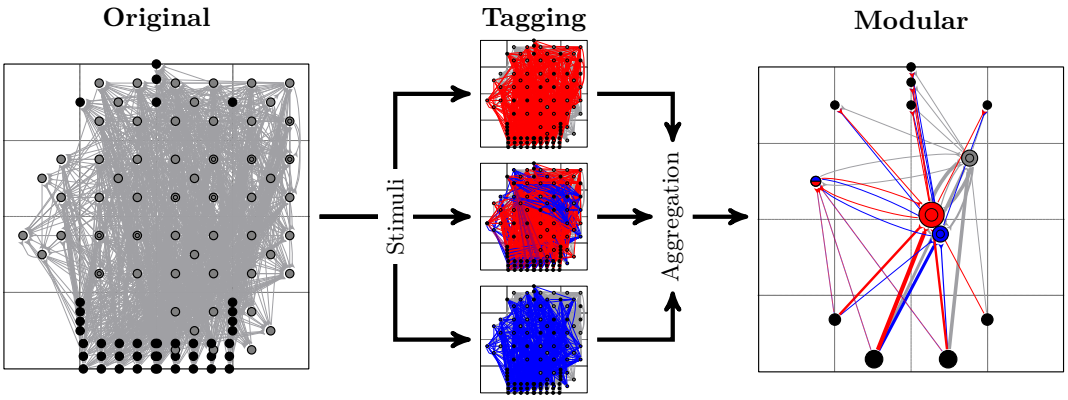


Figure 2: Producing a modular ANN based on individual neurons' activity patterns in canonical conditions. Left: a black box of hidden neurons and dense connections. Middle: we subject the creature to "simple" stimuli and note which neurons respond. Right: per-neuron response is aggregated, depending on the type of stimuli. For ease of reading the ANNs depicted here are displayed in a 2D substrate instead of the 3D ones used in this work.

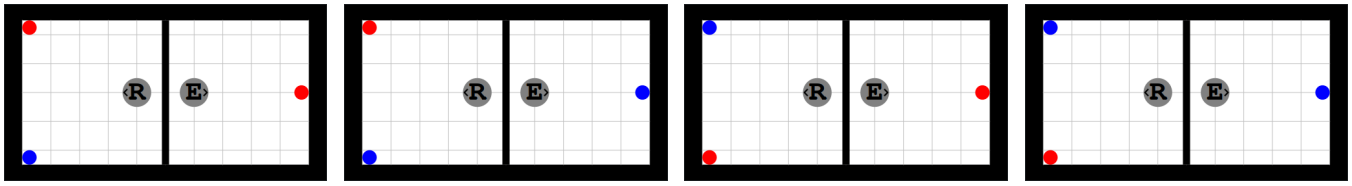
Virtual fMRI

To extract behavioral clusters, we use a procedure inspired by functional Magnetic Resonance Imaging (fMRI). In short, we subject an individual to a specific stimulus for pre-defined periods and note which neurons respond. We can then provide a functional mapping of the creature's brain according to the studied criteria thus streamlining subsequent investigations.

Indeed, whether through analysis of the neural pathways of the animal brain (Ledoux, 1998), the use of identified key cerebral regions to produce plausible artificial behavior (de Freitas et al., 2007; Delgado-Mata et al., 2007; Lotfi and Akbarzadeh-T., 2014) or mathematical approaches (Broekens et al., 2015), numerous methodologies have been devised to understand and mimic the biological brain. We argue, however, that all such approaches rely on the initial bias of our evolutionary history whereas artificial intelligence, explainable or not, could span from unexpected convergence of different factors. Especially with the highly scalable and geometry-aware ANNs produced by ES-HyperNEAT we can expect the resulting brains to exhibit natural patterns such as bilateral symmetry or partitions.

As introduced above, to accurately detect functional mapping between stimuli and neurons, we subject an individual to an alternating stimulus, in controlled conditions. In this experiment, we investigate modular clustering from two points of view: perceptions (vision/audition) and role (emitter/receiver). In all cases, the conditions are identical with only the type of applied stimulus changing between evaluations. Each individual is thus pinned (no movement allowed) to the center of an empty, wall-less arena. We expose the individual to the given stimulus for 2 simulated seconds (50 timesteps) followed by a relaxation of equal duration without said stimulus. This pattern is repeated 3 times resulting in 150 observations with and without the stimulus, thereby producing a dataset of differential activation. We then compare, for each neuron, their states depending on either the presence or absence of the studied stimulus via a one-sided Mann-Whitney test. All neurons that are found to have statistically different dynamics are then tagged as processing this stimulus, independently of their position or connectivity. When rendered with arbitrary colors, this allows for the visualization of functional areas (Figure 2).

Depending on the objectives of the given evaluation, the ANN is subjected to different classes of stimuli which are then aggregated. Subsequently, we generate modules from similar neurons depending on the associated flags. If these are null a default module is generated corresponding to the neurons unrelated to the specific scenario. Those who only responded to a single stimulus are rendered with a single color whereas those with multiple flags use a combination indicating their different functional responsibilities. To show the amount of aggregated items, module sizes are linearly scaled according to the number of neurons they abstract and inter-module connections are logarithmically scaled based on the number and weights of the underlying axons. Thanks to this procedure, we can observe the dynamics of an arbitrarily complex ANN via the prism of specific stimuli. The underlying network is left unchanged: the modular version merely provides the necessary perspective to monitor emergent higher-level structures and behavior. Furthermore, unlike alternative approaches, this methodology does not suffer from combinatorial explosion (Velez and Clune, 2016; Ghorbani and Zou, 2020) and does not require additional training (Csordás et al., 2020).



(a) 00 (b) 01 (c) 10 (d) 11

Figure 3: Environmental conditions for the evaluation of communication. The emitter (E) is placed in front of an object of a given color. The receiver (R) must reach the object of the same color on its side of the arena. To do so, E must vocally encode the color so that R can identify the correct target. Individuals are evaluated on all permutations (sub-tasks) of left/right-hand side configurations.

Experimental protocol

Unlike previous applications of these robots on foraging (Godin-Dubois et al., 2021) or competition (Godin-Dubois et al., 2023) tasks, the experiment described here focuses on collaboration in homogeneous pairs of robots. As illustrated on Figure 3, one robot (E) is placed on the right-hand side with a single-colored object while its counterpart (R) is always faced with two objects. The goal in each sub-tasks (a-d) is for the receiver to touch the object of the same color as the one seen by the emitter (positive reward). Naturally, the only way for this to happen is if the emitter transmits to its teammate the color of said object.

Evaluation of a single sub-task t is given by f_t as:

$$f_t(E,R) = \begin{cases} 1 + 0.5 \sum_{i \in \{E,R\}} energy(i) \\ \qquad\qquad \text{if R is right} \\ -1 \qquad\qquad \text{wrong} \\ 1 - dist(R) \quad \text{otherwise} \end{cases} \tag{1}$$

where $energy(i)$ is the normalized energy reserve of robot i at the end of the evaluation and $dist(i)$ is the distance between i and the correct object. Additionally, we reject invalid robots: those that either have no hidden neurons or that stay mute throughout the evaluation. The global fitness is then derived by averaging scores across all 4 sub-tasks. As both emitter and receiver share the same genotype, they also have identical ANNs requiring them to recognize, in some fashion, their role based on auditory/visual cues. However, to ensure that the emitter can see the target color, it is forced to remain immobile.

The *energy* term in Equation 1 is devised to provide a smooth gradient of improvement, as finding a solution that correctly reaches all objectives is only half the goal. A secondary objective is optimizing for energy efficiency which in this case includes axonal length and neural/motor/vocal activities, by order of importance. This is implemented by increasingly high metabolic costs to provide an incentive towards small, locally connected ANNs and parsimonious dialogs.

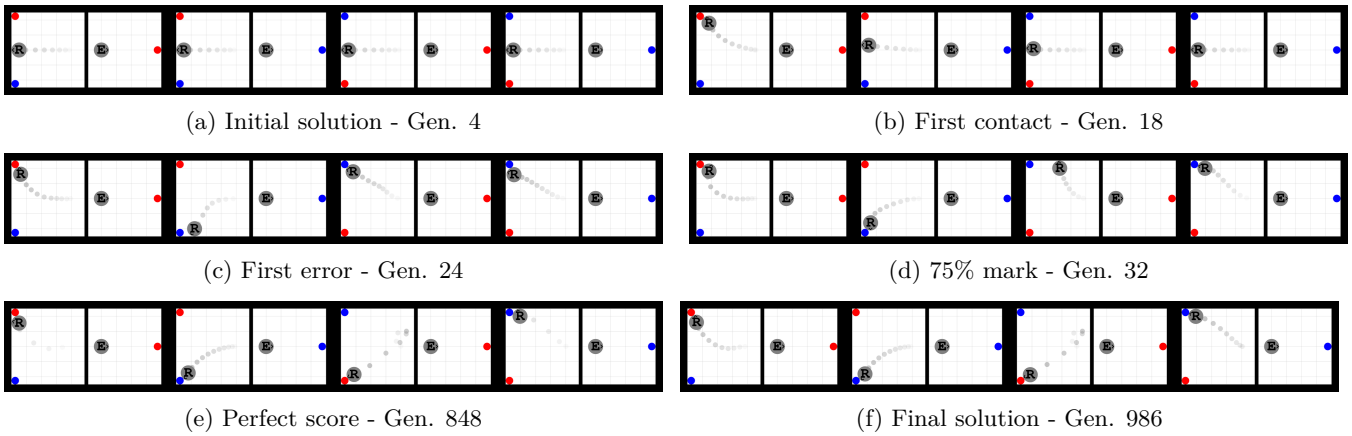
In this context, the expected high-level behavior is relatively straightforward: E should emit a color-specific vocal signal to indicate to R which color to look for. Indeed, both individuals only have a partial view of the game state requiring explicit sharing. Furthermore, we could expect, based on the energy requirements, that this communication would be one-directional with the receiver staying mute to preserve its reserves thereby increasing the joint fitness.

Populations of 200 individuals are evolved for 1000 generations with a Pareto-based tournament selection using the previously defined fitness (Equation 1) and a novelty metric. The latter takes into consideration energy expenditure for neural, motor and vocal activities as well as the mean and deviation of vocal patterns. One elite per criteria is preserved across generations and a total of 50 independent runs are conducted.

Population-level results

Overall, the evolved champions display efficient behavior for reaching the correct object. Thus, to put things into perspective, this section will address a few general points related to behavioral evolution, success rate and the ANNs.

A recurrent pattern, as illustrated in Figure 4, involves discovering independent solutions to each sub-task and integrating them, without loss of performance. In the first generations, the fittest individuals are those that get closer to the left side of the arena, thanks to the gradient provided by the last case of Equation 1. Building upon that state, individuals exhibit some divergent behavior with the general case still implying forward motion and an additional alternative: either going for one side of the arena or directly towards one object. The next step involves exploratory trajectory variations, occasionally resulting in erroneous collisions (e.g. Figure 4c), leading to an additional object being reached. Similarly, the third object is often found in early generations most efficiently when only minor adjustments of an existing trajectory are needed.



(a) Initial solution - Gen. 4

(b) First contact - Gen. 18

(c) First error - Gen. 24

(d) 75% mark - Gen. 32

(e) Perfect score - Gen. 848

(f) Final solution - Gen. 986

Figure 4: Cherry-picked example of strategy evolution. Starting from an unbiased state (a), the first object is reached by arbitrarily going one way (b). This solution is re-used in other cases, which can induce erroneous behavior (c). Three tasks are completed successfully as early as generation 32 (d) but it takes until generation 848 for a perfect scoring strategy to be found (e). Following changes are minimal until the end of the evolution (f).

However, the robots found it surprisingly hard to improve upon this state. For instance, in Figure 4e it took more than 800 generations to find a functional solution for all 4 permutations. During this time, the other trajectories remained by and large unchanged except for the speed at which the object is reached. In the latest generations, the pattern is repeated with only slight variations in speeds but not in the general trajectories.

On Figure 5, which generalizes these observations to the whole sample, we can see that the first two steps are completed extremely quickly with a median of 7.5 and and 13.5 generations for success rates of 25% and 50%, respectively. The same holds true for the third step as, although requiring more generations, most runs manage with less than a hundred offspring (median of 66). Conversely, reaching the perfect score is not only difficult for the previously shown lineage as we can see a very different trend to the other steps: the median is now at 479 with the fastest individual requiring about 50 generations. Furthermore, this only accounts for 46% of the replicates as the remaining half did not succeed in collecting objects in more than 3 of the sub-tasks.

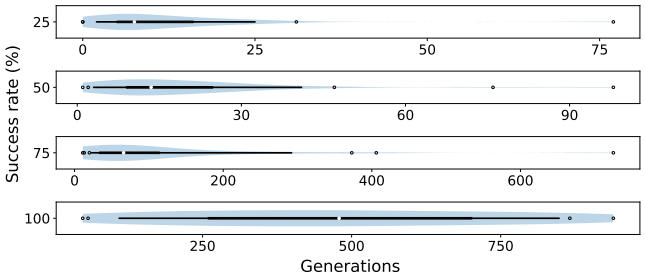


Figure 5: Violin plots of the number of generations needed to reach a specific success ratio. While it is almost trivially easy to reach the first two objects, getting to the third one proves more problematic. This is further magnified for the 100% success rate.

Neural networks

In terms of broad neural implementation of this communication task, one recurring feature is that the neural networks are of very limited size. Indeed, 76% of the sample is composed of networks between 4 and 8 neurons. Half of the remaining networks comprise more than 300 neurons each, interlinked by thousands of connections. As these ANNs have a relatively constant connectivity density, this results in a similar axonal distribution. Values are more dispersed as, unlike the neurons, they are not subjected to threshold effects.

As a consequence of ES-HyperNEAT's octree-based node discovery algorithm, neurons tend to be generated in bulk. Thus the simplest network is one composed of exactly eight neurons resulting from a single division of the octree's root: a pattern followed by 38% of this sample's ANNs. In practice, this can be related to the CPPNs themselves which are responsible for the NeuroEvolutionary process. While there was no relationship detected between the ANNs' size and the number of generations required to reach a perfect score, the opposite occurred for the CPPNs. Although with only moderate intensity, a correlation was found between the number of nodes/links contained in a CPPN with the number of generations taken to reach a perfect score. For node and links, this correlation was at 0.45 and 0.61, respectively, while only considering those runs that managed to get to all 4 objects. From this, we can gather that genetic bloating impeded generalization.

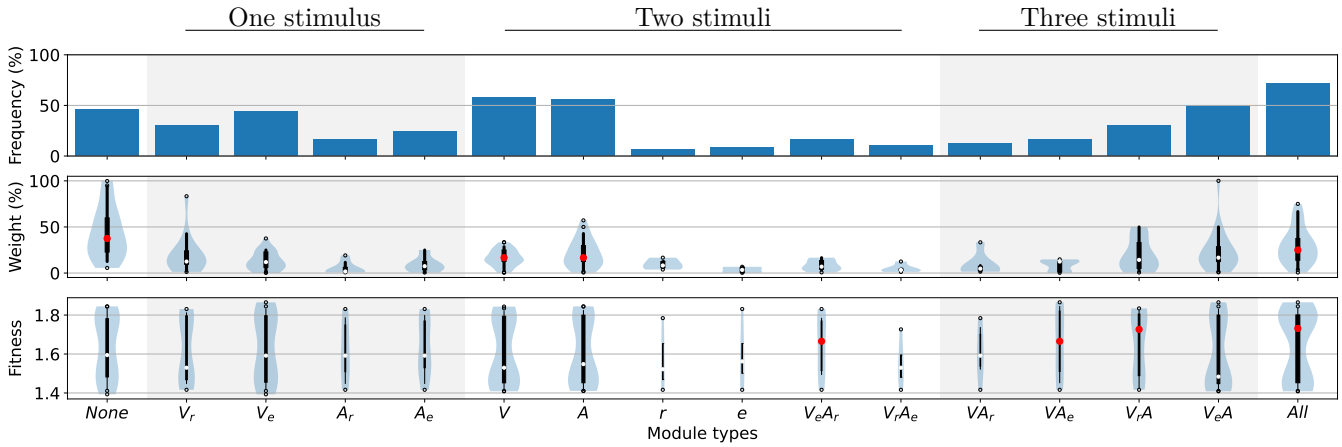


Figure 6: Module distributions by type. Top, the frequency of occurrence of a given module in the population. Middle, the proportion of the neural network devoted to said module (i.e. its dominance). Bottom, the fitness of individuals with a given module type. The highest four medians are highlighted in red and, whenever possible, labels represent the simplified name of a stimuli class (e.g. V instead of V_rV_e). Specialized modules V and A are both frequent and dominant while r and e are not. Generalist (Totipotent, All) modules are, additionally, significantly beneficial.

	Receiver		Emitter	
Visual	V_r^0	V_r^1	V_e^0	V_e^1
Auditive	A_r^{00} A_r^{10}	A_r^{01} A_r^{11}	A_e^{00} A_e^{10}	A_e^{01} A_e^{11}

Table 1: Stimuli set. Visual stimuli V_*^* consist of showing one of the agent's initial situation. Auditory stimuli A_*^i involve playback of a communication pattern in the corresponding scenario i.

Functional partitioning

The question addressed by this work lies not on whether the creatures succeeded at the task but rather on the *neural properties* that allowed them to do so. To investigate this point, we subject all final individuals to the virtual fMRI procedure with the set of stimuli enumerated in Table 1.

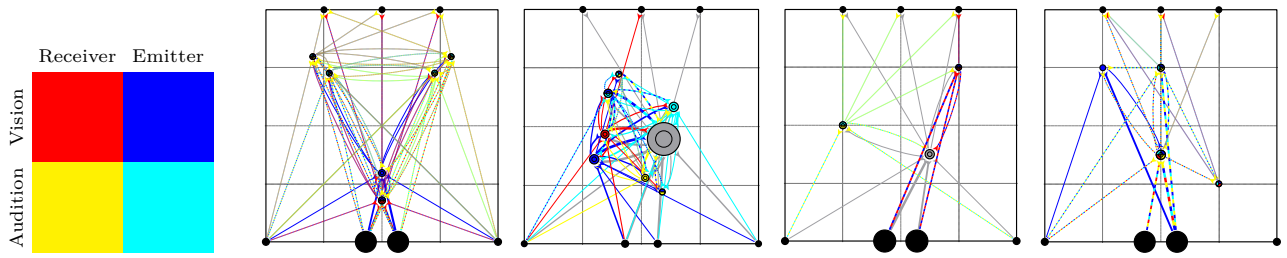
As they only possess two types of inputs, the set of available stimuli is quite straightforward. To ensure sufficient coverage of the various inputs the creatures might process, we enumerate every possibility. For the emitter's visual stimuli V_e^i this corresponds to presenting a single object either blue ($i = 0$) or red ($i = 1$). Similarly, for the receiver's stimuli, V_r^i references having two objects in one of the two permutations: red on the left and blue on the right for $i = 0$ and, conversely, blue on the left and red on the right for $i = 1$.

The case of auditory inputs is more complex as they not only differ from one run to the next but they also exhibit drastic variance both temporally and between teammates. To reduce the risk of processing artifacts while still providing sufficient stimulation to activate the relevant neurons we considered each sub-task independently. Thus the stimulus A_a^i relates to the auditory environment of agent a in sub-task i.

The problem now becomes one of extracting a meaningful audio sample from said sub-task, a trivial task when the teammate is continuously speaking but not so for the more parsimonious. To solve this we define the sample $S = (s_i)$ from the vocal output $V = (v_i)$ of its teammate as:

$$
\begin{aligned}
t_0 &= (i/v_i > 0 \wedge 0 \leq i < 50 \wedge \forall j < i, v_j = 0) \\
t_1 &= (i/v_i > 0 \wedge 0 \leq i < 50 \wedge \forall j > i, v_j = 0) \\
\hat{S} &= (v_i/t_0 \leq i \leq t_1) \cup (0) \\
S &= (\hat{S}_{i \bmod |\hat{S}|})_{0 \leq i \leq 50}
\end{aligned} \tag{2}
$$

which, informally, consists of looping the vocal output's first two seconds after removing silences at the beginning and end. In this manner, we ensure that all individuals can be evaluated with such stimuli: only a single instance exhibits a mute agent in one specific evaluation scenario. From this, we study the modules resulting from the union of similar elementary stimuli as shown in Table 1. For instance, module V_e corresponds to neurons found sensitive to the presence of a single blue or red object while A_r would be the neurons found responsive to auditory inputs similar to that produced by the emitter.



Module key (a) Symmetry (b) Jack-of-all-trades (c) Vision/Audition (d) Totipotency

Figure 7: Sample of modular ANNs from the best-performing individuals. (a,d) Numerous tri-modules. (c) Both visual and auditory specializations. (b) All stimulus-specific modules interacting with more generalists ones.

Specialization: In the context of this study, a specialized module is one that processes either a type of sense (vision, audition) or a specific role (emitter, receiver). As summarized in Figure 6, two of the most predominant types of modules are those related to visual (V) and auditory (A) processing indicating that evolved neural networks tend to partition the neurons between the tasks, validating hypothesis 1 (**H1**). At the same time, they only monopolize about a quarter of the neural capacity making sensory input an unobtrusive part of the network. Conversely, modules related to role (r, e) are infrequent, leading to the rejection of hypothesis 2 (**H2**). Over-specialized (stimulus-specific) neurons are also present in somewhat low numbers especially for audition.

Generalization: By opposition, a generalist module is one that process different types of inputs (here three or more). Surprisingly, the totipotent one (*All*) occurs in 72% of the population and is the most dominant of all but the neutral module. Given the relatively small size of the neural networks, these correspond to the other extremum of the spectrum where intricate re-use of the same neurons results in functional strategies.

One could conclude that, in this experimental setup, the choice between generalization and specialization went in favor of the former. Indeed, the totipotent module type is not only the most frequent and dominant but is also the only one which has clear benefits. On the last row of Figure 6, its median fitness is drastically higher than any other module type of comparable frequency. It is, however, unclear whether this is because module *All* is *inherently* beneficial or because it is more easily paired with other compatible modules.

To get a better picture of the kind of modular neural networks induced by this communication task, we present, in Figure 7, a sample taken from the best-performing individuals. The two outermost cases exhibit a totipotent module although with different topologies as 7a is almost fully connected while 7d relies on more heavily connected components. This is particularly visible for module VA_e (red, blue, aquamarine) which processes all inputs but the receiver's audition. Such partitioning is even more pronounced for 7c which relies on both types of sense-specific modules in combination with the *None* module. In this case, inputs are either routed to the dedicated neurons or the more general ones, contained by the neutral module.

Furthermore, there is evidence of emergent pathways for instance in the aforementioned peripherical module of 7d. While in most ANNs all neurons tend to be directly connected to the inputs, this module can only work on data pre-processed by other modules. This effect is drastically more present in 7b which combines multiple favorable characteristics such as its size (544 neurons) or the presence of all input-specific modules. More interestingly, it possesses the deepest neural network with a maximum of 5 connections between an input and an output. Taken together with the large neutral module, such features hint at more elaborate data processing capabilities not discovered with the current implementation of the VfMRI procedure.

Neural overlap

To further investigate how ANNs allocate neurons between the different senses and roles, we turn our attention to the underlying structures. More specifically, we want to measure how frequently a neuron is used for different types of input processing. To this end, we define three metrics Sep_V, Sep_A and Sep_R as follows.

$$SepV = \frac{\sum_n \mathbf{1}_{V_e^*}(n) \oplus \mathbf{1}_{V_r^*}(n)}{\sum_n \mathbf{1}_{V_e^*}(n) \vee \mathbf{1}_{V_r^*}(n)} \tag{3}$$

$$SepA = \frac{\sum_n \mathbf{1}_{A_e^*}(n) \oplus \mathbf{1}_{A_r^*}(n)}{\sum_n \mathbf{1}_{A_e^*}(n) \vee \mathbf{1}_{A_r^*}(n)} \tag{4}$$

$$SepR = \frac{\sum_n (\neg \vee_{t\in\{e,r\}} \mathbf{1}_{V_t^*}(n)) \wedge (\neg \vee_{t\in\{e,r\}} \mathbf{1}_{A_t^*}(n))}{\sum_n \vee_{s\in\{V_e^*,V_r^*,A_e^*,A_r^*\}} \mathbf{1}_s(n)} \tag{5}$$

where $\mathbf{1}_s(n)$ is the indicator function defining whether neuron n reacts to stimulus s. The first two measure whether either sense is implemented on different

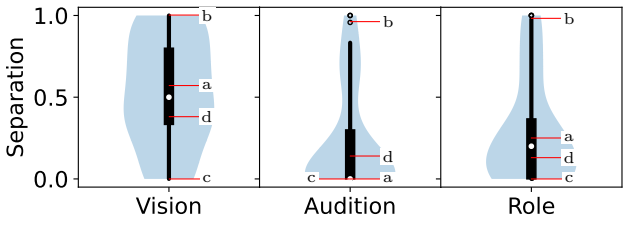


Figure 8: Distributions of the three separation metrics Sep_V, Sep_A and Sep_R. Individuals are homogeneous in terms of visual separation while there is a marked trend for re-use in auditory perception and role-specific processing. For comparison, individuals from Figure 7 are highlighted.

neurons for the different roles while the last one expands this measure to encompass both senses at the same time. This way we can study which strategy was preferred by evolution as summarized by Figure 8.

A cursory examination of the first case shows that individuals have not converged towards any side of the spectrum from which we can deduce that visual separation was globally neutral in terms of evolutionary advantage. Conversely, the auditory processing paints a drastically different picture with half the runs exhibiting no separation at all, i.e. they actively selected networks that re-used the same neurons for both agents' audition. This is consistent with previous observations that both emitter and receiver tend to fall into an extended dialog instead of devising short context-specific communication protocols. In a broader context, the neural implementation of role-specific tasks mostly falls on the side of reuse rather than specialized processing.

The modular ANNs from Figure 7 follow similar distributions: uniform for the vision and skewed towards lower values for audition and role. However, only the most extreme strategies (b,c) remain in a very narrow area. Indeed, the more moderate (a,d) have variable neural separation commensurate with the whole sample dynamics. As can be seen on the modular ANNs, they have separate vision-related modules and aggregated audition-related modules including the totipotent.

Conclusion

In this experiment, we used virtual robots to study whether NeuroEvolution would lead to specialization or generalization. These robots were embedded in a physical 2D environment which they perceived by simulated retina-based vision and frequential hearing. Locomotion and vocalization were their sole mode of interacting with said environment.

Placed in disjoint cells, two clones had to devise a communication scheme that allowed one agent to transmit color information to its teammate in order to reach the correct target. Broadly speaking, the experiment showed surprising results, with initial steps being completed much faster than expected, although tackling all four sub-tasks was a harder challenge. The resulting ANNs were of limited size, potentially as a runaway optimization response to the energy cost associated with neural and axonal activities.

While solving the task, these ANNs developed emergent topologies to process environmental data. Thanks to the VfMRI procedure, we extracted those structures to investigate whether different senses/roles were implemented on different neurons. The relationship between frequency, occupation and fitness highlighted diverging dynamics: visual and auditory processing were, indeed, specialized (**H1** accepted) while role-related modules were not (**H2** rejected). Complementarily, totipotent modules were found in large quantity, in individuals with generally higher fitness. By further investigating this dual tendency for specialization and generalization we found, through separation metrics, that auditory processing was markedly implemented on identical neurons. To a lesser extent, this also held true for the roles (emitter/receiver) while no clear trend were detected for visual inputs. Thus, this context highlights the balancing act between specialization and generalization with neither being the sole optimal solution.

In future work, a promising direction would be to force the receiver to be mute. While reducing the complexity of the system this should promote the emergence of more object-related, interpretable sounds, as shown by preliminary experiments. Additionally, a third color could also be incorporated into the experimental protocol to stimulate the emergence of more complex behavior and better understand the neural implementation of vocal communication.

Moreover, the observed emergence of "multi-layer" ANNs paves the way toward deeper input processing, allowing more complex environments to be tackled. At the same time, this would require extending the VfMRI procedure so that nth-order modules could be detected, for instance through iterative mapping of artificial inputs from (n-1)th-order modules. In this manner, Virtual functional Magnetic Resonance Imaging could be applied to an even broader range of tasks and neural representations including Deep Neural Networks and Embodied Robotics.

Acknowledgments As always, thanks to Alexandra Godin-Dubois for her invaluable editorial skills. The authors would also like to thank the members of the CI group (VU) for their informal contribution. This work was granted access to the HPC resources of CALMIP supercomputing center (allocation P16043).

References

Bacardit, J., Brownlee, A. E. I., Cagnoni, S., Iacca, G., McCall, J., and Walker, D. (2022). The intersection of evolutionary computation and explainable AI. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 1757–1762. ACM.

Broekens, J., Jacobs, E., and Jonker, C. M. (2015). A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, 27(3):215–233.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Campos, J. I. and Froese, T. (2019). From embodied interaction to compositional referential communication: A minimal agent-based model without dedicated communication channels. In *The 2019 Conference on Artificial Life*, pages 79–86. MIT Press.

Csordás, R., van Steenkiste, S., and Schmidhuber, J. (2020). Are Neural Nets Modular? Inspecting Functional Modularity Through Differentiable Weight Masks. *ICLR 2021 - 9th International Conference on Learning Representations*.

de Freitas, J. S., Imbert, R., and Queiroz, J. (2007). Modeling Emotion-Influenced Social Behavior for Intelligent Virtual Agents. pages 370–380.

Delgado-Mata, C., Martinez, J. I., Bee, S., Ruiz-Rodarte, R., and Aylett, R. (2007). On the Use of Virtual Animals with Artificial Fear in Virtual Environments. *New Generation Computing*, 25(2):145–169.

Gaier, A. and Ha, D. (2019). Weight Agnostic Neural Networks. *Advances in Neural Information Processing Systems*, 32.

Ghorbani, A. and Zou, J. (2020). Neuron Shapley: Discovering the Responsible Neurons. *Advances in Neural Information Processing Systems*, 2020-Decem.

Godin-Dubois, K., Cussat-Blanc, S., and Duthen, Y. (2021). Spontaneous modular NeuroEvolution arising from a life/dinner paradox. In *The 2021 Conference on Artificial Life*, page 95. MIT Press.

Godin-Dubois, K., Cussat-Blanc, S., and Duthen, Y. (2023). Explaining the Neuroevolution of Fighting Creatures Through Virtual fMRI. *Artificial Life*, 29(1):66–93.

Ito, T., Pilat, M. L., Suzuki, R., and Arita, T. (2013). Coevolutionary Dynamics Caused by Asymmetries in Predator-Prey and Morphology-Behavior Relationships. In *Ecal*, pages 439–445. MIT Press.

Kadam, S. and Vaidya, V. (2021). Cognitive Evaluation of Machine Learning Agents. *Cognitive Systems Research*, 66:100–121.

Kadish, D., Risi, S., and Beloff, L. (2019). An artificial life approach to studying niche differentiation in soundscape ecology. In *The 2019 Conference on Artificial Life*, pages 52–59. MIT Press.

Ledoux, J. (1998). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*.

Lotfi, E. and Akbarzadeh-T., M.-R. (2014). Practical emotional neural networks. *Neural Networks*, 59:61–72.

Mettler, H. D., Schmidt, M., Senn, W., Petrovici, M. A., and Jordan, J. (2021). Evolving neuronal plasticity rules using cartesian genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 285–286. ACM.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Olson, R., Hintze, A., Dyer, F., Moore, J., and Adami, C. (2016). Exploring the coevolution of predator and prey morphology and behavior. In *Proceedings of the Artificial Life Conference 2016*, pages 250–257. MIT Press.

Risi, S. and Stanley, K. O. (2012). An Enhanced Hypercube-Based Encoding for Evolving the Placement, Density, and Connectivity of Neurons. *Artificial Life*, 18(4):331–363.

Stanley, K. O. (2007). Compositional pattern producing networks: A novel abstraction of development. *Genetic Programming and Evolvable Machines*, 8(2):131–162.

Stanley, K. O., Clune, J., Lehman, J., and Miikkulainen, R. (2019). Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1(1):24–35.

Stanley, K. O., D'Ambrosio, D. B., and Gauci, J. (2009). A Hypercube-Based Encoding for Evolving Large-Scale Neural Networks. *Artificial Life*, 15(2):185–212.

Stanley, K. O. and Miikkulainen, R. (2002). Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, 10(2):99–127.

Tang, Y., Nguyen, D., and Ha, D. (2020). Neuroevolution of self-interpretable agents. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pages 414–424. ACM.

Treccani, C. (2020). The brain, the artificial neural network and the snake: why we see what we see. *AI & SOCIETY*.

Velez, R. and Clune, J. (2016). Identifying Core Functional Networks and Functional Modules within Artificial Neural Networks via Subsets Regression. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pages 181–188. ACM.

Yaman, A., Iacca, G., Mocanu, D. C., Coler, M., Fletcher, G., and Pechenizkiy, M. (2021). Evolving Plasticity for Autonomous Learning under Changing Environmental Conditions. *Evolutionary Computation*, 29(3):391–414.

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1).