Representational Similarity Analysis of Haemodynamic Responses during ATARI Video Gameplay using Deep Q Network Feature Regressors

Student:	Supervisor:	Supervisor:
Kevin Koschmieder	Marcel van Gerven	Umut Güçlü
Radboud University	Radboud University	Radboud University

October 2016

Abstract

In recent years, deep neural networks (*DNNs*) have come to dominate varied domains in the field of machine learning. Particularly novel are accomplishments in deep reinforcement learning and the successes of DNNs as feature models in neuroscientific studies, e.g. in the case of functional magnetic resonance imaging (*fMRI*) experiments investigating visual and auditory perception. This study represents the juncture of these two branches and aims to locate mechanisms of perceptual decision making by identifying the neural correlates of Deep Q Networks (*DQNs*).

In an fMRI experiment, 12 subjects played three conceptually different ATARI video games for which DQNs have been shown to achieve human-level performance. The Q-values and Hidden values of the DQN were used as feature regressors in a representational similarity analysis, analyzing correlation with the blood oxygen level dependant both at the scope of regions of interest and voxel searchlights.

The DQN generated features showed elevated correlations in occipital lobe. Actions caused heightened correlations in both visual- and motor-related areas. Furthermore, positive correlations were found in frontal lobe for Games. Albeit, statistical significance could not be established for these correlations.

Qualitatively, neural correlates were identified for all regressors in line with current neuroscientifc understanding. Potentially beneficial adjustments to the DQN and the study design were recognized, which might allow to fully exploit this new experimental paradigm in the future.

Key words: **Deep Q Network, functional magnetic resonance imaging, Q-learning, decision making, representational similarity analysis**

Introduction

On a daily basis, humans make countless decisions to navigate through their lives. Some decisions elude our consciousness because they are without (major) consequences or have become habitual, e.g. tying our shoes. Many decisions, however, incur rewards or punishments and thus require conscious deliberation of the possible choices. Such decisions vary in the time available to reach a decision, ranging from fast-paced reactions to long-term planning, and their complexity, i.e. how uncertain the reward-structure of a problem is. How humans (and animals) estimate, evaluate and develop behavioural policies is the focus of reinforcement learning (**RL**) neuroscience (Sutton and Barto, 1998).

RL is a framework of machine learning, in which agents optimize their behavioural policies with respect to rewards and punishments by interacting with the environment (Sutton and Barto, 1998). Several algorithms, e.g. temporal difference learning and Q-learning (Tesauro, 1995; Watkins and Dayan, 1992), have been proposed as models of how systems could learn optimal policies. These algorithms have been mathematically proven to converge, and to function in practical settings with artificial systems. They also led to many insights into human learning and decision-making by applying them as models both in psychological and neuroscientific investigations (Schultz, 2015).

Recently, Q-learning has been successfully combined with deep neural networks (**DNN**). Mnih et al. (2015) introduced the Deep Q Network (**DQN**), a convolutional neural network (**CNN**) that is capable of learning policies for various ATARI 2600 videogames en par with human performances. The DQN achieves this feat simply by playing the game, with no other information than the visual input, the possible actions, and the games' rewards. Since its introduction the DQN has gained popularity, inciting various improvements (Van Hasselt et al., 2015; Nair et al., 2015) and applications, e.g. in the field of autonomous driving (Schmidhuber, 2015).

Deep Learning itself has become a fundamental technique in the machine learning community in the last decade (LeCun et al., 2015), claiming state-of-the-art performance in many applications; reaching from image classification (Simonyan and Zisserman, 2014) over language translation (Bahdanau et al., 2014) to predicting financial trends (Längkvist et al., 2014), even excluding the aforementioned DQNs.

Recently, neuroscientific studies employed DNNs as models of cognitive mechanisms and representations in the human brain (Kriegeskorte, 2015), particularly in the domain of perceptual processing, as that is the area where DNNs originally garnered success (Krizhevsky et al., 2012) and which they are especially suited for (LeCun et al., 2015). Güçlü and van Gerven (2015a) found a hierarchical complexity in the various layers of CNNs, trained to classify images, similar to the gradient found in the ventral stream of visual processing, by mapping the unit activations to the blood oxygen level dependant (**BOLD**); an analogous discovery was made for the dorsal stream with a functional magnetic resonance imaging (**fMRI**) experiment using movies (Güçlü and van Gerven, 2015b). Other studies support the expressiveness of CNNs as a model for the human visual cortex (Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016). Beyond visual processing, Güçlü et al. (2016) also found neural correlates of learned features in auditory processing.

This study unified these two strands of developments by investigating the neural correlates of features as generated by self-trained DQNs in an fMRI experiment of continuous video gameplay. The recoreded BOLD responses were correlated with the Q-Values and Hidden Values of the DQN model using representational similarity analysis (RSA). Additional regressors, namely Games, Manual Actions, and In-Game Action, were introduced to provide control measures and validate this progressive experimental paradigm.

With this methodology we found qualitatively positive correlations both for the DQN features

and additional regressors, which are consistent with the current state of perception and action neuroscience. However, we were not able to match the DQN features to neural regions and mechanisms associated with reward estimation and evaluation, e.g. in striatal or frontal regions. Nonetheless, we were able to establish that this experimental paradigm has merit, and by identifying its pitfalls we were able to propose correctional measures for future applications.

Methods

Subjects

A total of 12 healthy subjects (age 21-29, 9 male and 3 female, 10 right-handed and 2 left-handed) participated in this study. Handedness was not a constraint in this study, as the tethyx joystick (**Fig. 1 (A)**) used in the experiment allowed for gameplay with each hand and the analysis did not distinguish between hemispheres. Video game experience and aptitude were not used as criteria in the selection of participants. However, a questionnaire was given to inquire these factors. (The questionnaire can be found in the **Supplementary Material**; it was ultimately not used in the analysis to divide the participant pool). The study was approved by the local ethics committee of Radboud University and the Donders Centre for Cognitive Neuroimaging.

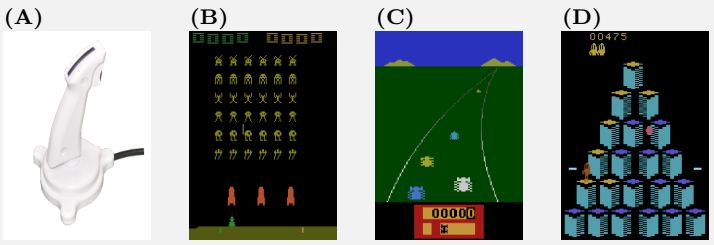

Figure 1: (A) Tethyx Fyber Optic Response Joystick: non-magnetic / non-electronic **(B) Space Invaders**: space shooter, player defends Earth against an army of approaching aliens (1980) **(C) Enduro**: racing game, player has to overtake as many cars as possible (1983) **(D) Q*bert**: puzzle game, player has to color all blocks by jumping on them while avoiding NPCs (1988)

Stimuli

Three ATARI 2600 video games were used in this study (**Fig. 1(B)-(D)**):

- ***Space Invaders***, a space shooter in which the player has to defend the Earth against a horde of attacking aliens;
- ***Enduro***, a racing game in which the player has to overtake a number of competitors each day;
- and ***Q*bert***, a puzzle game in which the player has to color a pyramid of boxes in a specific color by jumping on boxes under the threat of various computer opponents.

Presentation and gameplay was achieved by using the emulation capabilities of the Arcade Learning Environment (**ALE**). Games were chosen according to several criteria: human level performance of the DQN as determined by Mnih et al. (2015), reasonable number of controls, varying balance of reacting and planning behavior to succeed. Looking at the entirety of ATARI games, the range of planning is limited to short-term plans at the most, as the DQN struggles with games that require long-term strategies.

Experimental Design

Training The subjects were given short manuals to the games and instructed to read them. The manuals explained the objective of the game, described the controls and resulting actions, and noted any peculiarities of the game, e.g. the non-player characters and their abilities in the game *Q*bert*. (The manuals can be found in the **Supplementary Material**). Then they proceeded to play each game for the duration of 15 minutes in front of a notebook with the actual tethyx joystick used in the fMRI scanner. They were given additional two minutes per game in the fMRI scanner to get used to playing the games in supine position.

Experiment The subjects played three rounds of all three games, each game being played for six minutes without interruption. If the subject reached a *Game Over*-state, the game would automatically start anew. Independent of the respective situation, one run ended after six minutes. In all rounds, the order of games was fixed as seen in (**Fig. 1B-D**). We recorded the subject's videostream as frame-wise images, and their control inputs in each frame.

fMRI Data Acquisition

MRI data was collected at the Donders Centre for Cognitive Neuroimaging, Nijmegen. Functional BOLD volumes were acquired with a 3T-scanner (Prisma; Siemens) using a 32-channel head coil. A multi-band EPI sequence of $TR = 0.7$ s was used, i.e. for each run of six minutes, 530 fMRI volumes were collected. 64 slices with voxel size $2.4 \times 2.4 \times 2.4$ mm^3 were obtained. Additionally, a T1-weighted volume of higher resolution was collected for co-registration ($1.0 \times 1.0 \times 1.0$ mm^3).

fMRI Data Preprocessing

All preprocessing steps were performed with the SPM12 toolbox. First, the volumes were spatially re-aligned to the first volume. Then, the volumes were slice-time corrected to the first of the 64 slices. For the purposes of masking and group analysis, the volumes were translated into MNI-space. Finally, the individual runs were detrended. To account for the delay in haemodynamic response, the first and last 20 volumes of each run were discarded. For the Game-regressor, the analysis was performed over the entire session by concatenating the individual nine runs (after the aforementioned steps of detrending and discarding). In the case of the other regressors, the RSA was performed on single runs and the results averaged over games, subjects, and runs.

Behavioral Data Preprocessing

Four perceptual and behavioral features were considered. First, the different games were considered. Second, the player's actions were investigated. Here, we performed two analyses: one for the manual actions performed by the player, and the other for the available actions in each game. Possible manual actions were button press, horizontal joystick movement, vertical joystick movement, and the joystick resting position. An action was considered to be performed for its entire duration, not just its onset (or offset). Given the time-window of 0.7 s between TRs and assuming a stable frame-rate of 60 Hz, 42 actions were recorded per TR. For each TR, the ratio for all actions was computed and taken as the regressor.[1] The procedure was the same for the in-game actions. However, unlike Manual actions, each In-game action was considered separate, e.g. right and left movement

[1]The emulator cannot maintain this frame-rate perfectly, particularly at the start of the game the emulator needs a couple of seconds to stabilize the framerate. The description uses the ideal example. Computations for the analysis used the actual number of actions performed between each TR, not ideal approximations.

is bagged as horizontal movement for MANUAL ACTIONS, and action combinations (e.g. button + right) were considered positives for both button and horizontal movement, whereas they might be considered a unique action by game and DQN.

The other two features were the Q-VALUES and HIDDEN VALUES of the (un)trained DQN. In the following, the basics of this feature model and their extraction are explained. For further information, please consult (Mnih et al., 2015).

Deep Q Network (DQN)

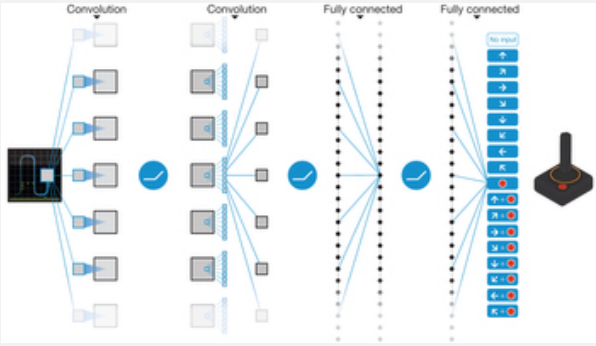

Figure 2: Schematic illustration of the convolutional neural network adopted from (Mnih et al., 2015). Further details to the architecture are given in the Methods. The input to the neural network is a concatenation of four grey-scale frames measuring 84×84 pixels. This is followed by three convolutional layers and one fully connected layer, and culminates in a fully-connected output layer with a single node for each valid action. Each hidden layer is nonlinearly rectified (meaning the activation x is bound in the positive: $\max(0, x)$).

The DQN is a CNN that is conventional in its architecture compared to standard CNNs used for classification purposes (Krizhevsky et al., 2012), but special in its way to learn. It is designed to learn a gameplay strategy for ATARI video games using end-to-end reinforcement learning without any prior knowledge except the number of a game's possible actions (this technically only lessens the computational burden, as the training would still work with all possible actions that the ATARI controller offers). The game's state-space, the effect of the actions, and its objective have to be inferred during the training procedure. Mnih et al. (2015) showed that this approach leads to human-level or better performance for a range of ATARI games.

Its architecture consists of three convolutional and one fully-connected layer (**Fig. 2**). It is fed the last 4 frames (resized to 84×84, grey-scale) as input. The output nodes represent the expected cumulative reward for each action. During evaluation (gameplay), the node with the highest expected reward is chosen by the DQN.

The convolutional neural network is designed to approximate the optimal action-value function, formalized as

$$Q^*(s, a) = \max_{\pi} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ... | s_t = s, a_t = a, \pi].$$

which is the maximum sum of rewards r_t discounted by γ at each time-step t, as defined by the Bellman equation. This maximum can be achieved with a behaviour policy $\pi = P(a|s)$, by responding to an observation (s) with an action (a) (Mnih et al., 2015).

During training, the DQN is iteratively updated according to the gradients computed with the loss function

$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim U(D)}[r + \gamma \max_{a'} Q(s', a'; \theta_i^-) - Q(s, a; \theta_i)]$$

in which γ is the discount factor determining the agent's horizon, θ_i are the DQN's parameters at iteration i computing the actual ouput, while θ_i^- are the network parameters used to compute the target at iteration i. The target network parameters θ_i^- are only updated with the Q-network

parameters (θ_i) at fixed intervals and are held constant between individual updates (Mnih et al., 2015). To this end, minibatches of 4-frame sequences are randomly selected from the algorithms memory replay ($(s, a, r, s') \sim U(D)$), which contains up to 1.000.000 of the last steps.

Model Architecture The DQN maps the sensory inputs, i.e. the game screen (in form of the last four frames), to Q-values for all available actions in any given game. The exact architecture is as follows. The input consists of an $84 \times 84 \times 4$ image, which is the concatenation of the last 4 images (re-sized and in terms of luminance). The first convolution layer contains 32 filter of 8×8 with stride 4 on the input layer and applies a rectifier non-linearity. The second convolutional layer contains 64 filter of 4×4 with stride 2 and applies a rectifier non-linearity. The third convolutional layer contains 64 filter of 3×3 with stride 1 and applies a rectifier. This is followed by a fully-connected linear layer of 512 rectifier units. The output layer is a fully-connected linear layer with a single node for each possible action (see **Fig. 2** and compare (Mnih et al., 2015)).

Training The DQN was trained for each game according to the parameters given in Mnih et al. (2015). For each game, networks the fully-trained network (200 epochs) and a randomly initialized, or untrained, network were taken to compute the Q-Values and Hidden Values used in the analysis. A python-theano-lasagne implementation was used to train and compute the DQNs (link to the original github repository and the repository containing adaptations can be found in the **Supplementary Material**).

Feature extraction The Q-Value feature represents the expected reward (activation value of output node) of the chosen action at each frame. Further approaches, e.g. comparing the best (determined by DQN) with the one actually chosen by the subject, or taking the whole output layer as a regressor, were discarded as the Q-values for all available actions were almost indistinguishable for most frames.

The Hidden Values feature contains all activation values of the 512 nodes in the fully-connected layer. As can be seen in the **Supplementary Material**, this feature vector is more distinctive than the Q-value feature scalar.

For each node, the forward pass (calculation of activation given the frame-wise video-stream and action log) through the DQN was performed at each frame. The resulting activation series of each node was convolved with a canonical haemodynamic function. Consequently, the regressors were sampled at the time of the TRs.

Representation Similarity Analysis (RSA)

RSA enables to relate information and representations from different modalities (Kriegeskorte et al., 2008); in our case it allows the correlation of the DQN as a computational model (and other regressors like actions) and the fMRI BOLD responses. This is achieved by comparing the patterns in so-called representational Dissimilarity Matrices (**DSMs**). For our purpose, the comparison is performed for a DSM of the respective feature regressor and one of the BOLD recording.

Regarding the DSM of the BOLD recording, there are different options to define the scope of voxels under consideration: according to regions of interest (**ROIs**) or voxel searchlights. Further details on these two modalities are provided in subsections below. Both approaches were used with the exception of the Game-regressor, as the searchlight RSA exceeded the computational ressources available.

A DSM is a square symmetric matrix measuring the representational distance between two states. There are various distance measures that have to be defined. One for the calculation of the

Table 1: Distance Matrix and Correlation Measures for Individual Regressors and Approaches

Regressor	Approach	Target DSM	Voxel DSM
Game	ROI	Binary	Pearson's r
Manual Action	ROI	Spearman's r	Spearman's r
Manual Action	Voxel Searchlight	Spearman's r	Pearson's r
In-Game Action	ROI	Spearman's r	Spearman's r
In-Game Action	Voxel Searchlight	Spearman's r	Pearson's r
Q-Value	ROI	Euclidean distance	Spearman's r
Q-Value	Voxel Searchlight	Euclidean distance	Pearson's r
Hidden Values	ROI	Spearman's r	Spearman's r
Hidden Values	Voxel Searchlight	Spearman's r	Pearson's r

distance within the feature or target DSM, another for the representational distance of the voxel DSM. There are a variety of distance measures we used: binary, euclidean distance, Pearson's r, and Spearman's r. Finally, a distance measure for correlating both DSM is necessary which was Pearson's r in all cases. Modality and regressor dependant use of distance measures is detailed in **Table 1**.

Playing video-games is a time-continuous task, unlike most decision-making related fMRI studies that apply designs with discrete decision events to accommodate the nature of fMRI, i.e. time-sparse collection of brain volumes. To accommodate the continuous nature of the task and the sparse recording with fMRI, the states are the time points of the fMRI pulse. (For visualizations of DSM for all regressors, please refer to the **Supplementary Material Fig. 1-5**.)

Region of Interest (ROI)

First, we performed RSA with respects to ROIs. This means, for the calculation of the DSM all voxels within a ROI are considered. The ROIs were defined with the SPM-extension *WFU pickatlas*, according to the IBASPM116 atlas (compare **Supplementary Material, Table 1**: whole atlas with abbreviations, acronyms were taken from Moore (1991).; **Supplementary Material, Fig. 8 & 9**: projections of respective ROIs).

Given the task of playing video games and the feature model at hand, not all ROIs are of relevance. While we performed RSA for all ROIs, the results reported are limited to those areas that were a-priori under investigation due to task and model, and those that showed surprising results (The remaining results can be found in the **Supplementary Material**). In the following, the important areas are detailed and justified.

The DQN, as described, is a machine learning model applying RL concepts to learn autonomous decision-making policies. The striatum has long been established as the center of human learning and decision-making (Schultz, 2015) and its subdivisions have been shown to perform a variety of tasks to these effects, e.g. caudate (**Cd**) and putamen (**Pu**) play roles in reward expectation and the computation of reward expectation errors (Haruno and Kawato, 2006). The pallidum (**Gp**) is involved in the selection of motor programs (Grillner et al., 2005). The thalamus (**Th**) serves as a focal point of information flow in the human cortex, where sensory information are relayed to other regions in the cortex (Sherman and Guillery, 2002).

Video-gaming is inherently a visual experience and the convolutional hierarchy of the DQN is inspired by human visual processing, which is known to be primarily performed in the occipital lobe. The first cortical structure in visual processing is V1, which is concentrated in the calcarine fissure (**ccs**) (Rockland and Ojima, 2003; Engel et al., 1994). Mid-level visual representations can be found in the cuneus (**Cun**) (Vanni et al., 2001). In the current literature, the distinction between individual areas of the occipital lobe is rarely done according to IBASPM116, but commonly visual

areas V1-V5 and higher order areas like MT. Nonetheless, the occipital areas are established in their role in visual processing, therefore it follows that the inferior occipital gyrus (**IOG**), middle occipital gyrus (**MOcG**), and superior occipital gyrus (**SOG**) be included in the results.

The parietal lobe is involved in multi-modal sensory integration (Lewis and Van Essen, 2000). It is considered the location of the dorsal stream of vision (Ungerleider and Haxby, 1994), i.e. processing spatial and motion information, and has been implicated in performing visuomotor transformations (Fogassi and Luppino, 2005). The precuneus (**PCu**) is implicated in visuo-spatial imagery and self-processing operations (Cavanna and Trimble, 2006). Spatial orientation as well as visuo-motor transformations are surmised to be functions of the superior parietal lobule (**SPL**) (Caminiti et al., 1996). The inferior parietal lobule (**IPL**) is also involved in spatial perception and visuomotor integration (Andersen, 2011). Lastly, the postcentral gyrus (**PoG**) is home of the primary somatosensory cortex, and thus the center of tactile processing (Kaas et al., 1979; Kurth et al., 2000).

The frontal lobe is involved in a variety of functions that relate to video gaming: reward, attention, memory, planning, and motivation (Miyake et al., 2000). These higher order functions are difficult to examine even with fMRI and there is still considerable research to be performed to understand the frontal lobe in its entirety. The superior frontal gyrus (**SFG**) contributes to working memory and spatial awareness (Du Boisgueheneuc et al., 2006). The inferior frontal gyrus (**BA45**) is surmised to play a role in attentional control (Hampshire et al., 2010), among other functions. The orbitofrontal cortex, consisting of superior frontal orbital gyrus (**OrGS**), middle frontal orbital gyrus (**OrGM**), and inferior frontal orbital gyrus (**BA47**), is indicated in emotion and reward-oriented decision-making (Bechara et al., 2000; Rolls, 2000). Primary motor cortex, as the name suggests, is heavily involved in the execution of movement and located in the precentral gyrus (**PrG**) (Hari et al., 1998; Karni et al., 1998). Also involved in movement, particularly in balance and planning, is the supplementary motor area (**SMA**) (Roland et al., 1980; Goldberg, 1985). Both motor areas play a huge role in finger movement (Shibasaki et al., 1993). Other frontal regions according to the atlas and subsequently investigated are superior frontal medial gyrus (**SFGm**), middle frontal gyrus (**MFG**), and inferior frontal gyrus pars opercularis (**BA44**).

Although anatomically these regions belong to two lobes, PrG, PoG, and SMA will sometimes be referred to as the motor-related areas in the remainder of this work.

Game worlds, as the real world, are comprised of objects, and the player has to identify these objects to successfully navigate these worlds, e.g. recognize the potential threats in the game Q*bert. The temporal lobe is considered to be part of the ventral stream of visual processing, e.g. object identification (Ungerleider and Haxby, 1994). The inferior (**ITG**) and middle (**MTG**) temporal gyri contribute to the visual processing of objects (Chao et al., 1999; Booth and Rolls, 1998). The fusiform gyrus (**FuG**) is commonly known as the 'face area', i.e. where facial features are processed and faces identified (Kanwisher et al., 1997), but this is not its only function. It is also surmised as the location of V4α, which is involved in color processing (Bartels and Zeki, 2000).

Searchlight

In a second step, we applied RSA on so-called spherical searchlights over the whole brain, i.e. voxel neighbourhoods of a specified diameter d (in our case $d = 3$). For each voxel in the brain, the sphere is determined, the dissimilarity between the sphere's voxels is calculated and then correlated to the DSM of the individual regressors (Kriegeskorte et al., 2006). After this, the results for the sphere were related to the respective ROIs for reporting.

Numerical results for all ROIs, both for the ROI and voxel searchlight-analysis, can be found in their entirety in the **Supplementary Material**. For the searchlight results, it has to be noted

that in this draft the maximal voxel for each region after averaging over all subjects and runs is reported. In the **Supplementary Material**, both the maximal searchlight as well as each ROI's average are listed.

Group Analysis

Given the nature of the task and the application of RSA in the manner described, i.e. with regressors that are so closely related (except for the GAME-regressor), the resulting correlations for both ROIs and voxel searchlights were (with very few exceptions) at least slightly positive. This means that standard statistical tests, which compare effects against a null-hypothesis were not applicable. The alternative approach of permutation testing, where the regressors are randomly mixed up for a large number of test-runs, was also not feasible because of the immense computation load of RSA for DSMs of this size.

As a result, no statistical measure could be applied. Thus, the following results are purely descriptive, comparing mean correlations over subjects, runs, and/or games for ROIs and voxel searchlights against average correlations over all regions or searchlights. Also to be noted, the thresholds and scale boundaries for the brain depictions were chosen arbitrarily. (Illustrations with optimized boundaries can be found in the **Supplementary Material**.)

Consequently, each result by itself does not imply a neural correlate. However, given the number of tests (region of interest and voxel searchlight, various regressors) and the existing literature a picture emerges that allows for qualitative interpretation.

Results

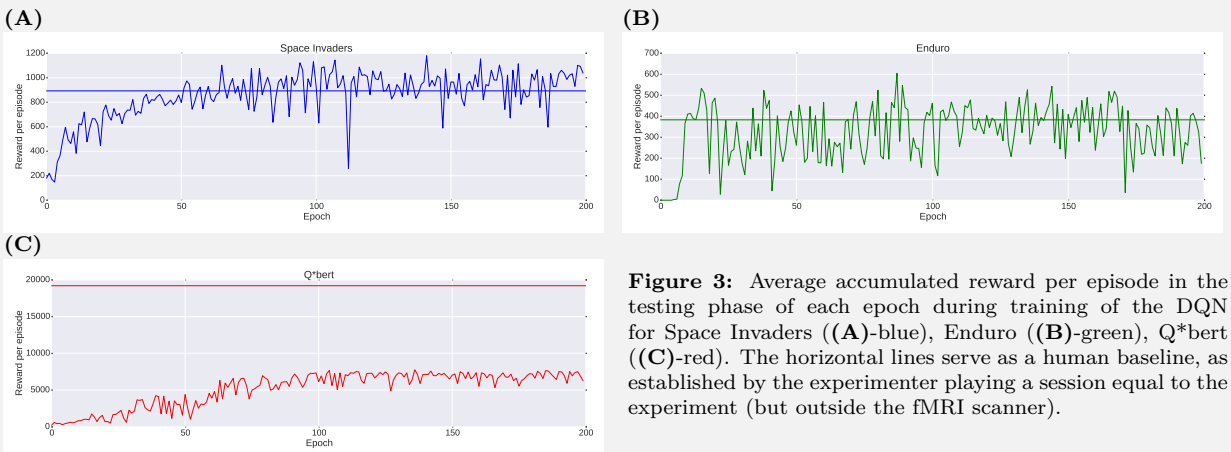

Figure 3: Average accumulated reward per episode in the testing phase of each epoch during training of the DQN for Space Invaders ((**A**)-blue), Enduro ((**B**)-green), Q*bert ((**C**)-red). The horizontal lines serve as a human baseline, as established by the experimenter playing a session equal to the experiment (but outside the fMRI scanner).

Deep Q Network Analysis

During training, the DQNs performance, measured in reward gained per episode[2], converged for *Space Invaders* and *Q*bert*, but not for *Enduro*, where it started to oscillated early in the training process and did not stabilize over the 200 epochs (see **Fig. 3(A)-(C)**). This happened for several random seeds.

[2] An episode is a single run until the Game Over state is reached.

In the original paper (Mnih et al., 2015), the trained DQNs resulted in human-level performance: the DQN achieved 121% for *Space Invaders*, 97% for *Enduro*, and 78% for *Q*bert*, compared to a baseline score accumulated by a professional human games tester. An average player was considered at a level of 75%. Please note, that the implementation used for this work is a different one from the one used in the original paper, despite implementing the same principle.

However, the reported numbers were not suitable in our case, as the subjects were not video game testers. For the purpose of comparison, we created a human baseline with training and playing time according to the durations reported in the methods section, by one of the experimenters (compared to the pool of participants, an adept player). It should be noted, that this baseline was established in a seated and not supine position, which eases the difficulty of play. **Fig 3** shows, that the DQN is on a similar level as the baseline for *Space Invaders* and *Enduro*, despite the lack of convergence for the latter. Although the DQN converged for *Q*bert*, the performance is significantly worse than the baseline.

Regressor: Games

RSA - Region of Interest

The analysis using the three different GAMES as regressor revealed a positive correlation (mean over all subjects) with major parts of the frontal lobe (see **Fig. 4**). Especially, the orbital gyri OrGs($r = 3.65e^{-3}$), OrGm ($r = 3.00e^{-3}$), BA47 ($r = 1.99e^{-3}$); compare **Supplementary Material, Table 2**) and the gyrus rectus ($r = 3.26e^{-3}$, not depicted) showed increased correlation. Additionally, IOG ($r = 1.99e^{-3}$) and ITG ($r = 1.73e^{-3}$) presented with above-average ($r_{\varnothing} = -0.20e^{-3}$) mean correlation with the games. There was also an unexpected positive correlation with the olfactory gyrus ($r = 0.90e^{-3}$, not depicted). Both, the ROIs in the parietal lobe and striatum presented exclusively with negative correlation.

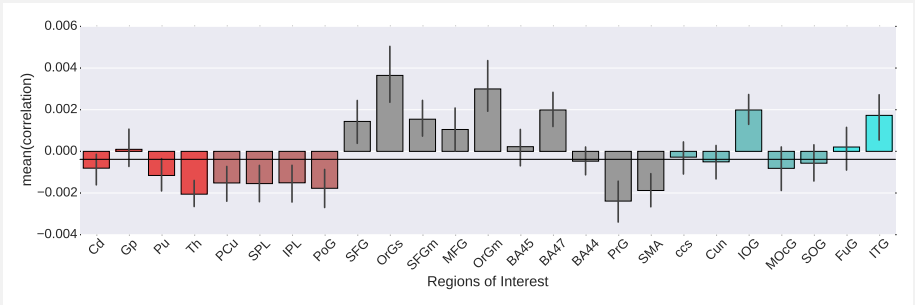

Figure 4: Mean (and sd) of Pearson's correlations given a target dissimilarity matrix based on GAMES (all runs concatenated) for selected regions of interest in striatum, parietal, frontal, occipital, and temporal lobe. The horizontal line shows the mean correlations of all regions of interest according to the IBASPM116 atlas.

Regressor: Actions

RSA - Region of Interest

MANUAL ACTIONS, i.e. button-presses and joystick movements, showed highest correlation (over subjects and runs) with ROIs in the occipital lobe (see **Fig. 5**). Over all games, there was increased correlation in ccs ($r = 6.21e^{-2}$), Cun ($r = 5.65e^{-2}$), IOG ($r = 4.31e^{-2}$), MoCG ($r = 5.01e^{-2}$), and SOG ($r = 5.84e^{-2}$) compared to an average of $r_{\varnothing} = 3.48e^{-2}$ (compare **Supplementary Material, Table 3**). This increase was particularly pronounced in the game *Q*bert*, and less noticeable for the game *Enduro*.

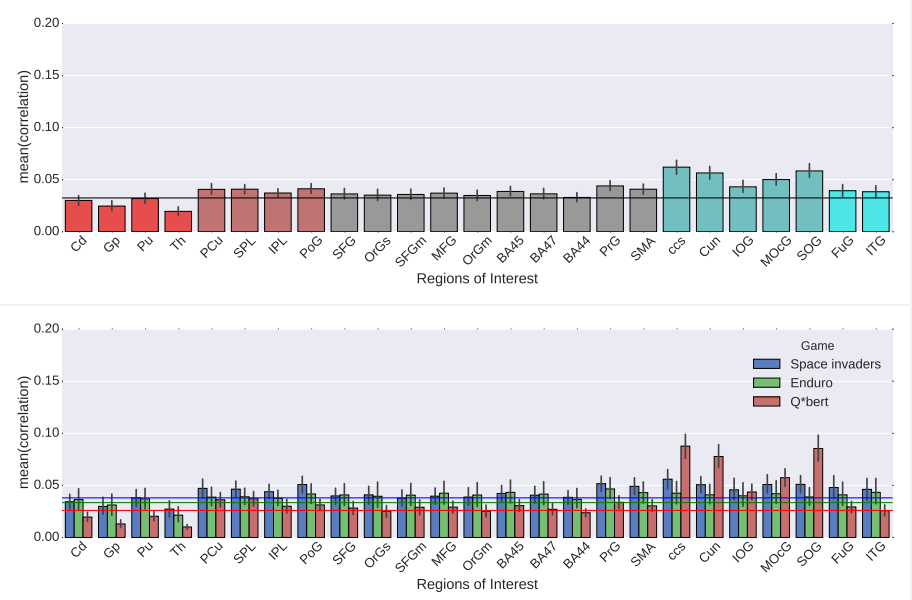

Figure 5: Mean (and sd) of Spearman's correlation given a target dissimilarity matrix (Spearman correlation) based on MANUAL ACTIONS, i.e. button-press, horizontal and vertical joystick movement and no action, (calculated for individual runs) for selected regions of interest in striatum, parietal, frontal, occipital, and temporal lobe. The upper graph shows the mean for all games and runs, the lower graph shows the mean for each game respectively. The horizontal line shows the mean correlations of all regions of interest according to the IBASPM116 atlas.

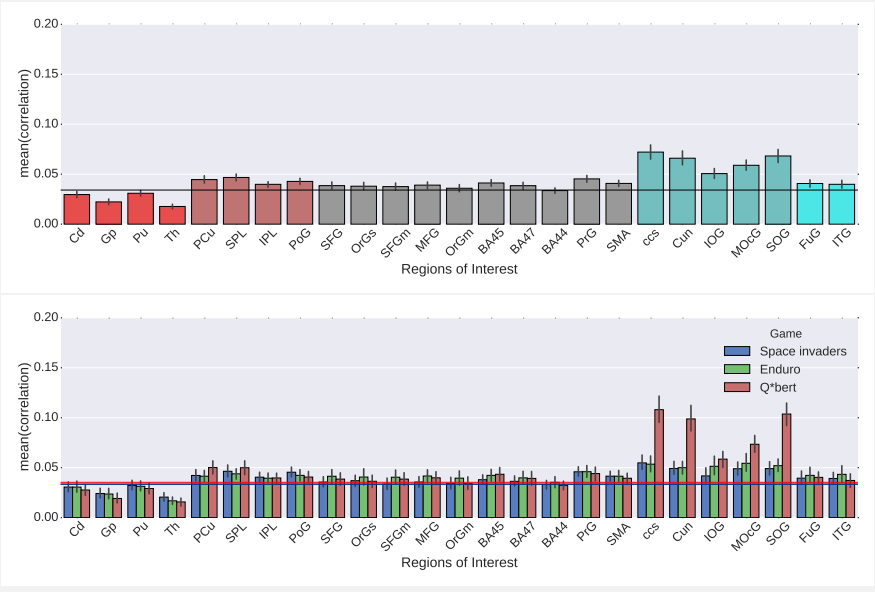

Figure 6: Mean (and sd) of Spearman's correlation given a target dissimilarity matrix (Spearman correlation) based on IN-GAME ACTIONS (calculated for individual runs) for selected regions of interest in striatum, parietal, frontal, occipital, and temporal lobe. The upper graph shows the mean for all games and runs, the lower graph shows the mean for each game respectively. The horizontal line shows the mean correlations of all regions of interest according to the IBASPM116 atlas.

ROIs in the frontal lobe were close to the average mean correlation. Slightly above mean correlation were the parietal and temporal ROIs. Mostly below the average mean correlation was the striatum. Against initial intuition, the motor-related areas PrG ($r = 4.13e^{-2}$), PoG ($r = 4.40e^{-2}$), and SMA ($r = 4.09e^{-2}$) did not present the strongest correlation with this regressor, but were consistently above-average for all games respectively.

A similar picture emerged with the analysis of the IN-GAME ACTIONS, i.e. the various actions and/or combinations available to the player in the individual games (see **Fig. 6**). The striatum was below-average, the regions of the frontal lobe average and the temporal regions slightly above. Moreover, the correlations of the motor-related areas were similar for the two action interpretations. Stronger correlations were observed for the parietal regions PCu ($r = 4.46e^{-2}$) and SPL ($r = 4.67e^{-2}$) (compare **Supplementary Material, Table 4**), and overall for the ROIs in the occipital

lobe.

RSA - Searchlight

The voxel searchlight analysis was coherent with the results from the ROI analysis. For both types of action regressors, MANUAL (see **Fig. 7**) and IN-GAME (see **Fig. 8**), the visual areas showed the highest correlation. Additionally, the correlations in the motor-related PrG and PoG (compare **Supplementary Material, Table 7 & 8**) appeared more pronounced than in the ROI analysis. Also, the differences between the respective game became apparent. The correlations were similar between perception- and action-oriented ROIs for Space Invaders, and presented a focus on action or perception for Enduro and Q*bert, respectively.

Please note that the boundaries of the figure scales were determined by the average and highest mean correlation for all games; the lower threshold being set at 120% of the average mean and the upper threshold being set at 90% of the highest mean correlation. (for surface projections with scales individually fixed for each game, see **Supplementary Material, Fig. 10 & 11**)

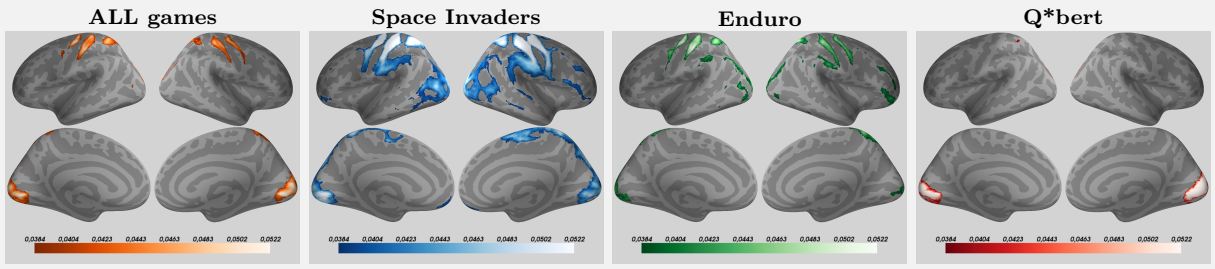

Figure 7: Mean of Pearson's correlation given a target dissimilarity matrix (Spearman correlation) based on MANUAL ACTIONS, i.e. button-press, horizontal and vertical joystick movement and no action, (calculated for individual runs) for average surface projections of full-brain searchlights measuring voxel size 3. From left to right, the color represents the game(s) used: orange for *ALL games*, blue for *Space Invaders*, green for *Enduro*, and red for *Q*bert*.

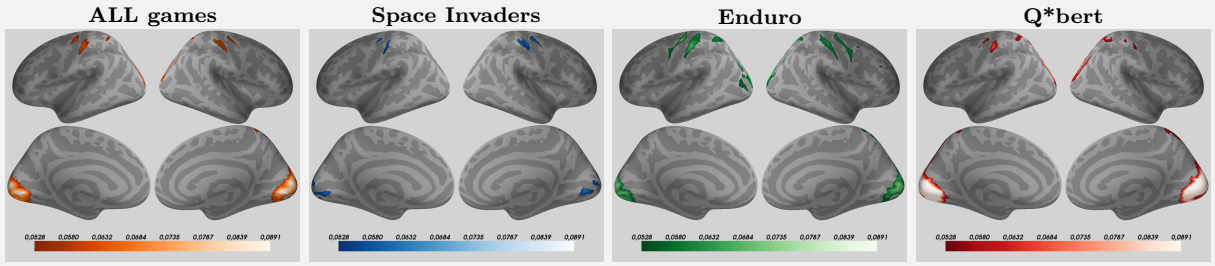

Figure 8: Mean of Pearson's correlation given a target dissimilarity matrix (Spearman correlation) based on IN-GAME ACTIONS (calculated for individual runs) for average surface projections of full-brain searchlights measuring voxel size 3. From left to right, the color represents the game(s) used: orange for *ALL games*, blue for *Space Invaders*, green for *Enduro*, and red for *Q*bert*.

Regressor: Q-Values

RSA - Region of Interest

For the Q-VALUE regressor (see **Fig. 9**), i.e. the DQN calculated expected rewards, there was no ROI that elevated strongly over the average mean correlation of $r_{\varnothing} = 4.72e^{-2}$ for all games (compare **Supplementary Material, Table 5**), although all non-striatal regions that we focused on did at least match it. Looking at individual games, the correlations were more distinctive. For *Space*

Invaders, the occipital regions, e.g. ccs ($r = 9.53e^{-2}$) and Cun ($r = 8.07e^{-2}$), showed the highest correlation, the parietal and motor regions elevated correlations, and the frontal and temporal regions close to average correlation ($r_{\varnothing} = 5.86e^{-2}$). A similar distribution, though less pronounced, was found for *Q*bert*. The distribution of correlations was different for *Enduro*, where the frontal ROIs showed the highest correlation. Nevertheless, no major spike above average correlation ($r_{\varnothing} = 4.94e^{-2}$) was detected.

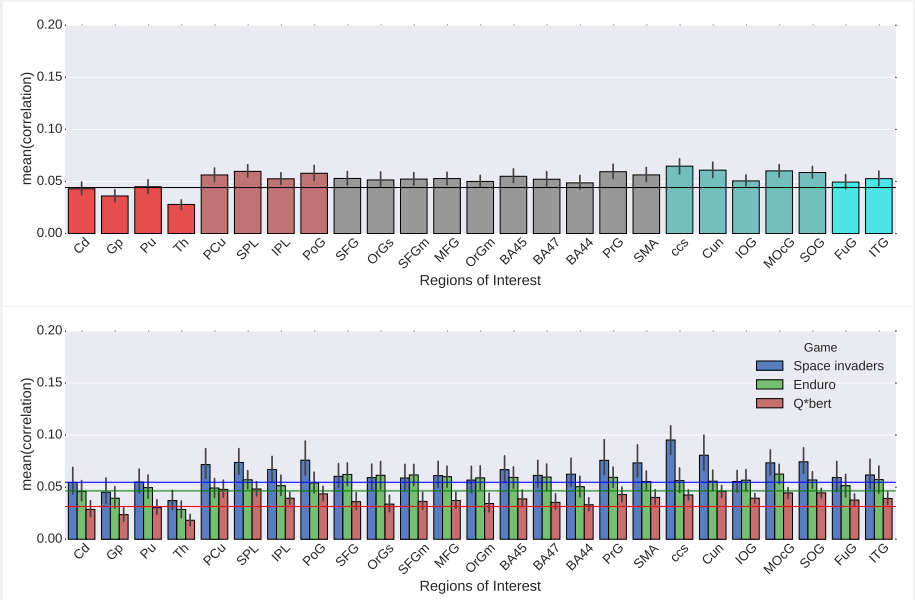

Figure 9: Mean (and sd) of Spearman's correlation given a target dissimilarity matrix (euclidean distance) based on Q-VALUES for the chosen action as determined by the DQN (calculated for individual runs) for selected regions of interest in striatum, parietal, frontal, occipital, and temporal lobe. The upper graph shows the mean for all games and runs, the lower graph shows the mean for each game respectively. The horizontal line shows the mean correlations of all regions of interest according to the IBASPM116 atlas.

RSA - Searchlight

The searchlight results for the Q-VALUES also showed the highest correlations in the occipital lobe (see **Fig. 10**), with maxima of $r = 7.12e^{-2}$ for ccs, $r = 6.95e^{-2}$ for Cun, and up to $r = 6.38e^{-2}$ in the occpital gyri over all games (compare **Supplementary Material, Table 9 & 10**). Elevation over the average was strongest in these regions during *Space Invaders*, and similar for *Enduro* and *Q*bert*.

A second, but less pronounced, peak over average ($r_{\varnothing} = 4.92e^{-2}$) in correlation could be found in motor-related areas PrG ($r = 5.44e^{-2}$), PoG ($r = 5.27e^{-2}$), SMA ($r = 5.31e^{-2}$) and the parietal regions IPL ($r = 5.51^{-2}$) and SPL ($r = 5.42^{-2}$). Out of the general focus of this analysis, there was also a strong correlation in the lingual gyrus ($r = 6.34^{-2}$).

For the randomly initialized DQN, i.e. the untrained network, heightened correlations in the visual and motor regions could also be found. Their correlations were overall closer to the average correlation than for the fully-trained network.

Fig. 10 depicts no noteworthy correlations for *Q*bert*, neither for the trained nor untrained network, and a surprisingly large surface area with higher correlations for *Enduro*. This is due to a considerable difference in average mean correlations over all searchlights (please refer to **Table 9/10** in the **Supplementary Material** for the exact numbers) despite uniform scales in this paper.

Please note that the boundaries of the figure scales is determined by the average and highest mean correlation for all games of the fully-trained network. The lower threshold being set at 120% of the average mean and the upper threshold being set at 90% of the highest mean corre-

lation. (for surface projections with scales individually fixed for each game and network-type, see **Supplementary Material, Fig. 12**)

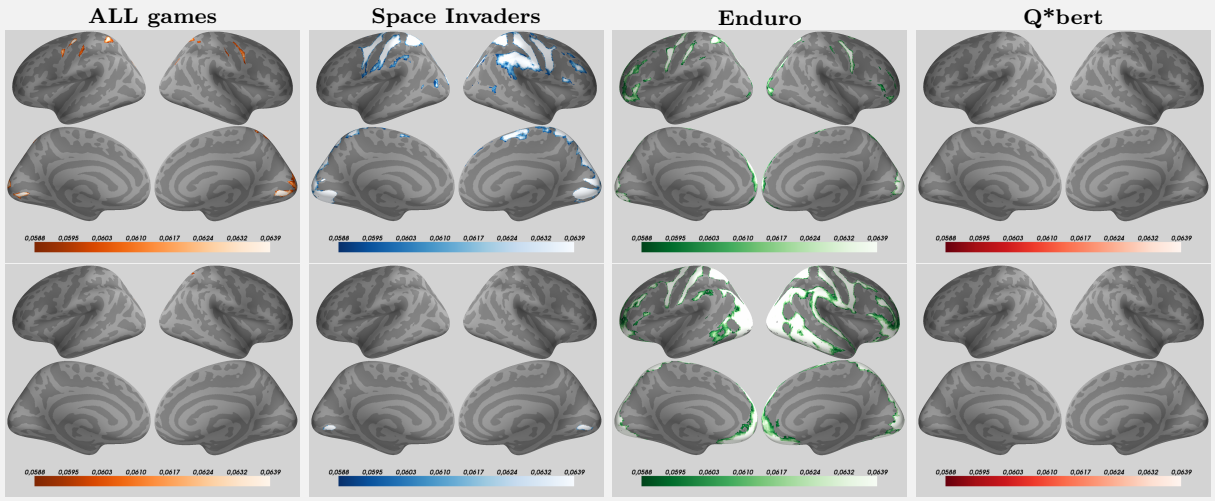

Figure 10: Mean of Pearson's correlation given a target dissimilarity matrix (euclidean distance) based on Q-VALUES for the chosen action as determined by the DQN (calculated for individual runs) for average surface projections of full-brain searchlights measuring voxel size 3. The upper row depicts the mean for a network trained for 200 epochs, the lower row depicts the mean for a randomly initialized network. From left to right, the color represents the game(s) used: orange for *ALL games*, blue for *Space Invaders*, green for *Enduro*, and red for *Q*bert*.

Regressor: Hidden Values

RSA - Region of Interest

For the HIDDEN VALUE regressor, above average ($r_{\varnothing} = 5.49e^{-2}$, compare **Supplementary Material, Table 6**) correlations could be found for all non-striatal regions we focused on, with strong elevations for the occipital and parietal regions (see **Fig. 11**), e.g. ccs ($r = 9.30e^{-2}$) and SOG ($r = 8.94e^{-2}$), and PCU ($r = 7.35e^{-2}$) and SPL ($r = 7.42e^{-2}$). On the level of individual games, these peaks in the distribution showed for *Space Invaders* and *Q*bert*. For those two games, the motor regions showed also elevation over average. For *Enduro*, the distribution was different, with overall weaker correlations in occipital and parietal lobe, but in comparison stronger correlations in the frontal lobe.

RSA - Searchlight

Unlike the ROI analysis, strong elevations in correlations only appeared in the occipital regions (see **Fig. 12**). The average correlations of all regions was $r_{\varnothing} = 9.72e^{-2}$ (compare **Supplementary Material, Table 11 & 12**); ccs presented with $r = 16.48e^{-2}$, Cun with $r = 16.06e^{-2}$, and SOG with $r = 15.46e^{-2}$.

While there was some elevation in correlation for the parietal regions in *Enduro* and *Q*bert*, overall the parietal and motor-related regions were closer to the average than for all other regressors (excluding the GAME-regressor). Again, the lingual gyrus showed a heightened correlation with $r = 12.66e^{-2}$. This increase over average is very localized.

For the untrained net, the increased correlation could also be localized in the occipital regions. As was the case with the analysis of Q-values, the increase in correlation was less pronounced for these regions.

Please note that the boundaries of the figure scales is determined by the average and highest mean correlation for all games of the fully-trained network. The lower threshold being set at 120% of the average mean and the upper threshold being set at 90% of the highest mean correlation. (for surface projections with scales individually fixed for each game and network-type, see **Supplementary Material, Fig. 13**)

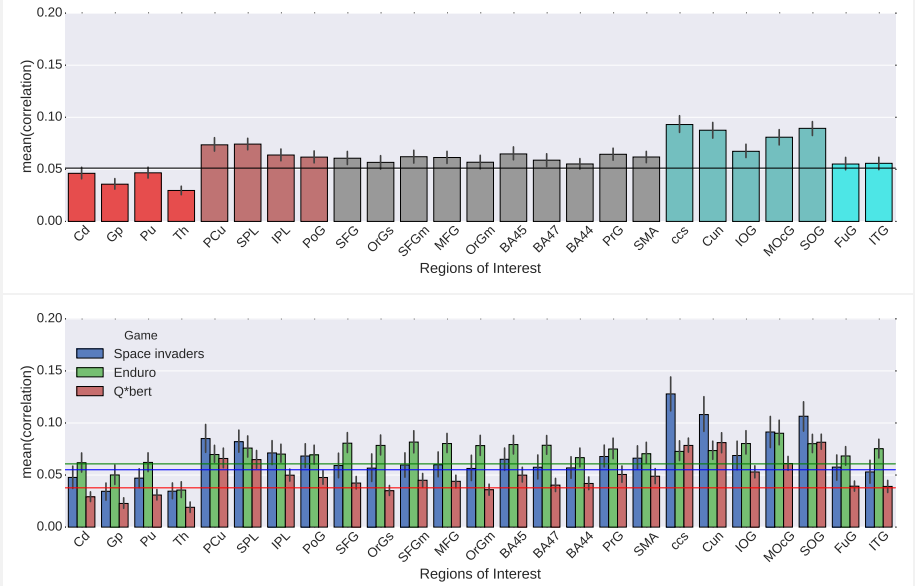

Figure 11: Mean (and sd) of Spearman's correlation given a target dissimilarity matrix (Spearman correlation) based on HIDDEN VALUES, i.e. the 512 unit values of the hidden layer of the DQN (calculated for individual runs) for selected regions of interest in striatum, parietal, frontal, occipital, and temporal lobe. The upper graph shows the mean for all games and runs, the lower graph shows the mean for each game respectively. The horizontal line shows the mean correlations of all regions of interest according to the IBASPM116 atlas.

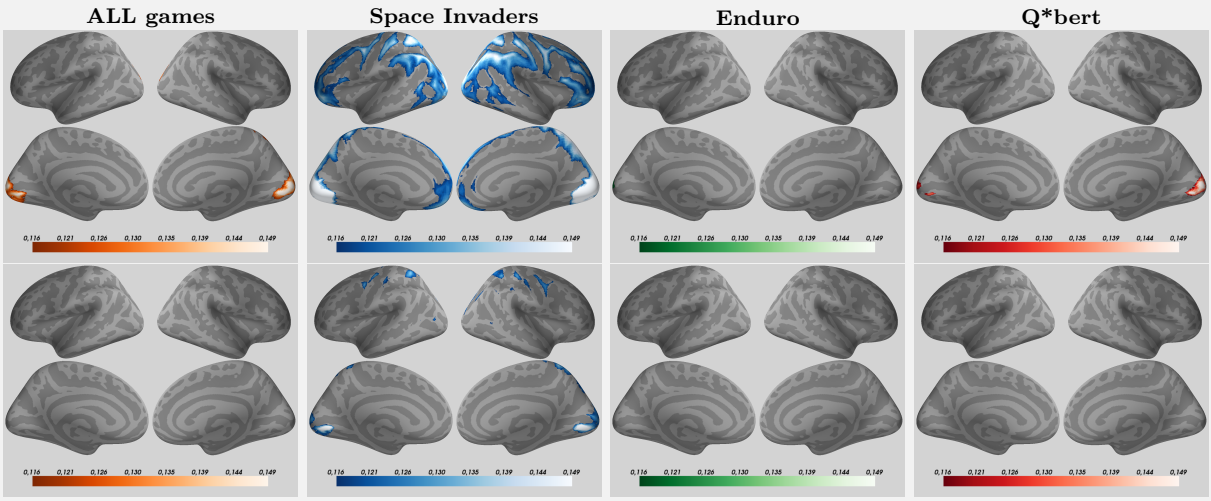

Figure 12: Mean of Pearson's correlation given a target dissimilarity matrix (Spearman correlation) based on HIDDEN VALUES, i.e. the 512 unit values of the hidden layer of the DQN (calculated for individual runs) for average surface projections of full-brain searchlights measuring voxel size 3. The upper row depicts the mean for a network trained for 200 epochs, the lower row depicts the mean for a randomly initialized network. From left to right, the color represents the game(s) used: orange for *ALL games*, blue for *Space Invaders*, green for *Enduro*, and red for *Q*bert*.

Discussion

Neural correlates Despite the problems establishing statistical significance with the described experimental and analytical design, which we will detail further below, we were able to find positive or elevated correlations which are coherent with the existing literature in the field of neuroscience, particularly with studies involving DNNs, for the features of the DQN and other regressors relevant to video-gaming.

For the GAME-regressor, i.e. the distinction of games, positive correlations were found in the fronto-orbital gyri, as well as the IOG and ITG, which contribute to the ventral stream of visual processing in the human brain (Reddy and Kanwisher, 2006). The latter is an expected finding, given that the three games are visually different (compare **Fig. 1(B)-(D)**. The cause for the positive correlation in the frontal lobe can only be guessed considering its vast functions. Nonetheless, the result is not surprising since the three games have varying demands in terms of attention and planning (reactive v planning), as well as having different reward patterns no matter if we follow the game's internal patterns or potential schemes human players might use.

The action regressors, both MANUAL ACTIONS and IN-GAME ACTIONS, correlated with visual and motor-related areas. The motion-related correlations were expected, though against intuition, the correlations were weaker for motion than for visual processing. This can be explained with the small action feature space. Ultimately, all motions were performed with hand or finger and were rather close and often overlapping. This makes it harder to establish correlates compared to a case where one would use different body parts. The correlations with visual mechanisms and representations can be explained two-fold. First, in ATARI games similar perceptual states lead often to the same action, e.g. in *Enduro*, a car appearing on the right side will often be outmaneuvered with a motion to the left. Second, each action leads to perceptually very similar outcomes in ATARI videogames, even though they only affect a part of the screen, e.g. movement or button fire in *Space Invaders*. Since the BOLD signal cannot be recorded at a speed that allows the distinction between the visuals preceding and following an action, this adds to this effect.

Q-VALUES, as determined by the DQN, were not the best option for a RSA feature value, as they are a scalar and show little variance over a run. Consequently, RSA has to correlate minuscule variations in the BOLD-signal. Nonetheless, correlations could be found in occipital lobe and to a lesser degree in parietal and motor areas. This is not quite the aspired result, as it was hoped to find correlates of expected reward in the striatum or prefrontal cortex. However, given the DQN model the result is understandable. Mnih et al. (2015) showed with a t-SNE analysis that the game screen could be clustered both according to their visuals as well as their associated reward, meaning similar visual states share similar expected rewards.

Unlike the Q-VALUES, the HIDDEN VALUES from the fully-connected layer of the DQN consisted of 512 values and thus, allowed for more distinctive DSMs. As a result, the effects are more focused in the occipital lobe. These representations in the visual cortex are coherent with the notion of CNNs and previous results from neuroscientific inquiries of DNNs (Güçlü and van Gerven, 2015a; Kriegeskorte, 2015; Yamins and DiCarlo, 2016).

Deemed very positive is the fact that the correlations are stronger in the visual cortex both for the Q-VALUES and the HIDDEN VALUES, when comparing the fully-trained and untrained networks. This is a good indicator that the DQN actually learned meaningful representations that are similar to the representations of human players.

Is fMRI data acquisition sufficiently fast and accurate to adopt a video game experiment? Videogames (like other games) come in many different shapes and forms, but tend to require quick decision-making; multiple decisions per second are typical and Mnih et al. (2015) assumed a rate

of six decision per second for the DQN as humanly feasible. The games used in this study are no exception. As such, fMRI is a seemingly unsuitable approach. First of all, the BOLD signal measured is not completely understood, very slow to follow the cognitive process that caused it ($\sim$ 6 s) and that delay may be location specific (Huettel et al., 2004). Additionally, current fMRI scanners can record volumes at frequencies beyond to 2 Hz (TR= 0.5 s), but the possible resolution and signal-to-noise-ratio are in a trade-off with this acquisition speed. As mentioned, the imaging protocol used here had a TR of 0.7 s which under the given assumption of decision-speed means up to four different action signals muddled into a single volume.

This is not to say that fMRI cannot produce meaningful results for fast-paced stimuli like movies (Güçlü and van Gerven, 2015b). However, the interaction with the videogame makes for a more engaging and complex cognitive process than watching a movie or listening to a piece of music. Consequently, it becomes more difficult to isolate meaningful representations and mechanisms in the brain.

In any event, there are no alternative imaging tools that could be applied in a more meaningful way than fMRI. The only faster maging tools are EEG and MEG. EEG lacks the spatial resolution (Aine, 1994), and MEG is too sensitive to motion artifacts to allow for a normal videogaming experience (Junghöfer et al., 2000).

Also worth mentioning is, that studies targeting the striatum usually use a multi-band-multi-echo protocol , instead of a simple multi-band protocol, as it allows for a better signal in deeper cortical regions (Huettel et al., 2004). However, multi-band-multi-echo protocols have a TR of $\sim$ 2 s, which makes their temporal resolution even worse for video-gaming than the applied protocol and explains why decision-making neuroscience normally applies very controlled experimental designs with large pauses. As this study was more exploratory in nature and looked at the full-brain, the choice for the multi-band protocol was uncontested. Nonetheless, the issue of protocol should be considered when looking at the resulting correlations for striatal regions. The lack of elevated correlation may entirely be caused by an incompatibility of the DQN model, but may also be partially due to the unfavorable imaging protocol.

Imaging acquisition modalities, among them fMRI, will continue to improve allowing for faster protocols with better signal-to-noise ratio. This alone will aid researchers in adopting similar experimental design. There are other possibilities to address this issue even short-term. The ALE could be easily adapted to be limited to a different frame-rate, thus decreasing the speed of the stimuli to e.g. half (30 Hz). With this the gaming experience would be adjusted to better match the recording speed.

RSA - effect sizes, t-tests and other issues The results of the RSA presented with two major issues: small effect sizes, and their incompatibility with secondary analysis like t-tests. Looking at previous studies applying RSA (Kriegeskorte, 2015; Devereux et al., 2013; Laakso and Cottrell, 2000), the effect sizes were usually on the scale of $r = e^{-1}$. For this study, the results were on the scale of $r = e^{-3}$ for the GAME regressor and $r = e^{-2}$ for all other regressors, and as such they appear to be too small to contain any meaning. This is a misconception; mathematically these effect sizes can be explained by the unprecedented size of the dissimilarity matrices used in this approach. The difference between the regressors' results are by a factor of 10, which is approximately the difference in size of their DSMs. For the GAME regressor, the DSM measures 4410×4410 and 490×490 for all other regressors.

For secondary statistics, e.g. the t-test, (apart from the GAME regressor) all regions for the ROI analysis and almost all voxel searchlights presented with almost exclusively positive correlations. Consequently, the whole brain were determined to be significant. While video-gaming certainly

presents a task, that involves many regions to some extent, such a significance obviously cannot hold true for every regressor. As such, these tests hold no meaning and present a major problem for this approach. At the root of this issue is the question why basically every resulting correlation is positive.

There is no clear answer to this, but there are several aspects that might contribute to this effect. Given the continuous task, the RSA might suffer from auto-correlations in the BOLD signal, which could be worsened by the sparsity of the feature space for the regressors, e.g. the action space is very limited and often overlapping and the Q-VALUES subject to miniscule changes resulting in overall unvaried DSMs. Furthermore, video-gaming is a task that encompasses almost the entire cortex in an interactive fashion - visual perception, movement, reaction, decision-making, and so forth. Unlike in standard fMRI studies, where a task is usually investigated in isolation and consequently allows for a stronger signal of the relevant brain area. Also, given the low feature and action space of ATARI games, visuals, actions, and rewards are often connected, which might cause additional auto-correlation.

Deep Q Network - a suitable model of the human mind playing video games? Deep Neural Networks have gained some popularity in neuroscience in recent years, and this study aimed at showing that the positive results would also hold for a neural network that is an actor, not just a classifier. Despite the positive results, there are issues with the DQN as a model of human decision making.

To simplify gradient computations, the learning algorithm clipped all rewards to either $+1$, 0, or -1. This is not congruent with human reward valuations, for humans the amount of reward matters and there are indicators that learning functions differently for rewards and punishments (Delgado et al., 2000). In general, the reward schemes that are implemented in the games most likely differ from the internal reward scales the human players, particularly novices who do not pay much attention to the reward meter on the screen, apply. For example, avoiding one of the opponents in *Q*bert* is not awarded in the direct way of points (but is indirectly valued, as the expected cumulative reward is higher after training), but might be a very meaningful event for a player, because they managed to escape from danger.

Furthermore, compared to other neuroscientific studies using deep neural networks (Güçlü and van Gerven, 2015b; Güçlü et al., 2016; Yamins and DiCarlo, 2016), the DQN has a rather shallow architecture with just three convolutional and one fully-connected layer. The DNNs used in (Güçlü and van Gerven, 2015a) have at least thrice the depth. Given that the visual complexity of ATARI games is rather low (especially compared to natural images or today's videogames) and the necessary training time (several days per game), this is understandable, but this also means that the representations the DQN can learn are comparably lower in complexity.

DNNs are currently a major focus of the machine learning community. Subsequently, there are weekly extensions and improvements to this field as a whole. This shows how adaptable of the deep learning framework is to all kinds of problems. For a study of human video gaming, there are several paths one could take to make the model more human. First, the DQN is not limited by a reaction time like human players are. If this constraint were added, the model might develop policies that are closer to their human counterparts. Second, the DQN, like many other RL algorithms, suffers from the exploration v exploitation dilemma (Mnih et al., 2015; Sutton et al., 1999). It already contains the experience replay and iterative update to counter this, but mainly for the purpose of optimization. Our interest in further exploration is that the calculated expected rewards should be more distinctive, which is difficult for the algorithm to achieve if it starts to mainly experience its successful strategy. Alternatively, the beginning of the DQN training could be exchanged for

an instruction session, in which the DQN basically watches human gameplay to learn. This would require crowd sourcing of such data and was beyond the scope of this work. But this would improve training time, and could ultimately result in policies that resemble human gameplay.

Game differences The correlates of the MANUAL ACTIONS and IN-GAME ACTIONS regressors were stronger than the DQN correlates for *Enduro*. A possible explanation is that the DQN, despite learning a semi-successful policy, was not able to converge for this game. Accordingly, the representations and action values learned were not fully developed and optimized.

Another remarkable result was that action-related correlations for *Q*bert* were more poignant in the occipital regions than for the other games. Additionally, *Q*bert* was the only game to illicit notable correlations in the parietal lobe, not just for the action-related regressors but also for the DQN computed features. The cause for this is difficult to pin-point. We can only surmise that the visuo-motor transformations that are believed to take place in SPL (Caminiti et al., 1996) are more stable for *Q*bert* than for the other two games.

In this paper, we have treated the average mean correlation over one or all games as a baseline for determining whether correlations in regions or searchlights can be qualitatively described as meaningful, at least with some confidence. Problematic with these comparisons is, that within one feature regressor the deviation between these average mean correlations can be quite substantial, e.g. $r_{\varnothing} = 13.17e^{-2}$ for *Space Invaders* and $r_{\varnothing} = 8.06e^{-2}$ for *Q*bert* (see **Supplementary Material, Table 11**). There is no constant order to these deviations, and there is also a case where these values are very similar (see **Supplementary Material, Table 4**). Given the issues with statistical significance tests, the qualitative comparison remained the only viable option in this case, and the results are consistent with themselves and the literature. However, this aspect has to be considered in this discussion.

Summary and Outlook This study establishes a promising precedence for the combination of DQNs and videogaming in neuroscientific investigations. Despite the fact that statistical significance could not be established for all regressors, the positive findings are coherent with the current understanding of visual processing and the recent findings of neural correlates of DNNs. However, the primary objective of finding neural correlates of reward expectation, evaluation and decision-making could not be achieved. Reasons for this could be the DQN-model, the analysis with RSA, and the experimental design. We were able to identify issues with all of the cogs of this study and suggested appropriate measures for future studies attempting to take a similar approach.

References

Aine, C. J. (1994). A conceptual overview and critique of functional neuroimaging techniques in humans: I. mri/fmri and pet. *Critical reviews in neurobiology*, 9(2-3):229–309.

Andersen, R. A. (2011). Inferior parietal lobule function in spatial perception and visuomotor integration. *Comprehensive Physiology*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bartels, A. and Zeki, S. (2000). The architecture of the colour centre in the human visual brain: new results and a review. *European Journal of Neuroscience*, 12(1):172–193.

Bechara, A., Damasio, H., and Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, 10(3):295–307.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279.

Booth, M. and Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex*, 8(6):510–523.

Caminiti, R., Ferraina, S., and Johnson, P. B. (1996). The sources of visual information to the primate frontal lobe: a novel role for the superior parietal lobule. *Cerebral Cortex*, 6(3):319–328.

Cavanna, A. E. and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583.

Chao, L. L., Haxby, J. V., and Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature neuroscience*, 2(10):913–919.

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D., and Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of neurophysiology*, 84(6):3072–3077.

Devereux, B. J., Clarke, A., Marouchos, A., and Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *The Journal of Neuroscience*, 33(48):18906–18916.

Du Boisgueheneuc, F., Levy, R., Volle, E., Seassau, M., Duffau, H., Kinkingnehun, S., Samson, Y., Zhang, S., and Dubois, B. (2006). Functions of the left superior frontal gyrus in humans: a lesion study. *Brain*, 129(12):3315–3328.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., and Shadlen, M. N. (1994). fmri of human visual cortex. *Nature*.

Fogassi, L. and Luppino, G. (2005). Motor functions of the parietal lobe. *Current opinion in neurobiology*, 15(6):626–631.

Goldberg, G. (1985). Supplementary motor area structure and function: review and hypotheses. *Behavioral and brain Sciences*, 8(04):567–588.

Grillner, S., Hellgren, J., Menard, A., Saitoh, K., and Wikström, M. A. (2005). Mechanisms for selection of basic motor programs–roles for the striatum and pallidum. *Trends in neurosciences*, 28(7):364–370.

Güçlü, U., Thielen, J., Hanke, M., and van Gerven, M. A. (2016). Brains on beats. *arXiv preprint arXiv:1606.02627*.

Güçlü, U. and van Gerven, M. A. (2015a). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, 35(27):10005–10014.

Güçlü, U. and van Gerven, M. A. (2015b). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *NeuroImage*.

Hampshire, A., Chamberlain, S. R., Monti, M. M., Duncan, J., and Owen, A. M. (2010). The role of the right inferior frontal gyrus: inhibition and attentional control. *Neuroimage*, 50(3):1313–1319.

Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., and Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proceedings of the National Academy of Sciences*, 95(25):15061–15065.

Haruno, M. and Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of neurophysiology*, 95(2):948–959.

Huettel, S. A., Song, A. W., and McCarthy, G. (2004). *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland.

Junghöfer, M., Elbert, T., Tucker, D. M., and Rockstroh, B. (2000). Statistical control of artifacts in dense array eeg/meg studies. *Psychophysiology*, 37(04):523–532.

Kaas, J. H., Nelson, R. J., Sur, M., Lin, C.-S., and Merzenich, M. M. (1979). Multiple representations of the body within the primary somatosensory cortex of primates. *Science*, 204(4392):521–523.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, 17(11):4302–4311.

Karni, A., Meyer, G., Rey-Hipolito, C., Jezzard, P., Adams, M. M., Turner, R., and Ungerleider, L. G. (1998). The acquisition of skilled motor performance: fast and slow experience-driven changes in primary motor cortex. *Proceedings of the National Academy of Sciences*, 95(3):861–868.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, 10(11):e1003915.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Kurth, R., Villringer, K., Curio, G., Wolf, K.-J., Krause, T., Repenthin, J., Schwiemann, J., Deuchert, M., and Villringer, A. (2000). fmri shows multiple somatotopic digit representations in human primary somatosensory cortex. *Neuroreport*, 11(7):1487–1491.

Laakso, A. and Cottrell, G. (2000). Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76.

Längkvist, M., Karlsson, L., and Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42:11–24.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lewis, J. W. and Van Essen, D. C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. *Journal of Comparative Neurology*, 428(1):112–137.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive psychology*, 41(1):49–100.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Moore, E. (1991). BrainInfo. `http://braininfo.rprc.washington.edu/Default.aspx`. [Online; accessed 20-November-2016].

Nair, A., Srinivasan, P., Blackwell, S., Alcicek, C., Fearon, R., De Maria, A., Panneershelvam, V., Suleyman, M., Beattie, C., Petersen, S., et al. (2015). Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*.

Oosterhof, N. N., Connolly, A. C., and Haxby, J. V. (2016). Cosmomvpa: multi-modal multivariate pattern analysis of neuroimaging data in matlab/gnu octave. *bioRxiv*, page 047118.

Reddy, L. and Kanwisher, N. (2006). Coding of visual objects in the ventral stream. *Current opinion in neurobiology*, 16(4):408–414.

Rockland, K. S. and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, 50(1):19–26.

Roland, P. E., Larsen, B., Lassen, N., and Skinhoj, E. (1980). Supplementary motor area and other cortical areas in organization of voluntary movements in man. *Journal of neurophysiology*, 43(1):118–136.

Rolls, E. T. (2000). The orbitofrontal cortex and reward. *Cerebral cortex*, 10(3):284–294.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological reviews*, 95(3):853–951.

Sherman, S. M. and Guillery, R. (2002). The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1428):1695–1708.

Shibasaki, H., Sadato, N., Lyshkow, H., Yonekura, Y., Honda, M., Nagamine, T., Suwazono, S., Magata, Y., Ikeda, A., Miyazaki, M., et al. (1993). Both primary motor cortex and supplementary motor area play an important role in complex finger movement. *Brain*, 116(6):1387–1398.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sprague, N. (2015). Theano-based implementation of deep q-learning. `https://github.com/spragunr/deep_q_rl`.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y., et al. (1999). Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063. Citeseer.

Tesauro, G. (1995). Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68.

Ungerleider, L. G. and Haxby, J. V. (1994). 'what'and 'where'in the human brain. *Current opinion in neurobiology*, 4(2):157–165.

Van Hasselt, H., Guez, A., and Silver, D. (2015). Deep reinforcement learning with double q-learning. *CoRR, abs/1509.06461*.

Vanni, S., Tanskanen, T., Seppä, M., Uutela, K., and Hari, R. (2001). Coinciding early activation of the human primary visual cortex and anteromedial cuneus. *Proceedings of the National Academy of Sciences*, 98(5):2776–2780.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.