Extensive characterization of human brain activity under naturalistic stimulations for developing individual artificial neuronal models

Experimental Protocol

Researcher :

Pierre Bellec, Ph. D. :

Centre de recherche, IUGM
4545 chemin Queen-Mary
Montréal, Québec
Canada H3W 1W5
Courriel: pierre.bellec@criugm.qc.ca

Co-Investigators :

- Sylvie Belleville, Ph. D., Département de psychologie, Université de Montréal, (514) 340-3540, poste 4767.
- Simona Brambati, Ph. D., Département de psychologie, Université de Montréal, (514) 340-3540, poste 4147.
- Julien Cohen-Adad, Ph. D., Département de génie électrique, Université de Montréal, (514) 340-4711, poste 2264.
- Adrian Fuente, Ph. D., École d'orthophonie et d'audiologie, Université de Montréal, (514) 343-6111, poste 37180.
- Jean-Pierre Gagné, Ph. D., École d'orthophonie et d'audiologie, Université de Montréal, (514) 340-3540, poste 4125.
- Karim Jerbi, Ph. D., Département de psychologie, Université de Montréal, (514) 343-6111, poste 29549.
- Pierre Rainville, Ph. D., Faculté de Médecine dentaire, Université de Montréal, (514) 340-3540, poste 4145.
- Pierre Orban, Ph. D., Département de psychiatrie, Université de Montréal, (514) 251-4015, poste 3553.

Summary and objectives

Understanding how intelligence works is one of the key frontiers of human knowledge. Recent impressive advances on this front have been made in a field of computer science called Artificial Intelligence (AI). Researchers in AI are not interested in understanding the human brain as such, but rather to apply simple yet powerful learning techniques to solve real-world challenges such as face recognition, image annotation, or playing games. An important trend in AI is called deep learning: artificial deep neural networks (DNNs) featuring many layers of computational units, which are able to match or even exceed human level performance on specialized tasks. Current DNNs architectures however tend to be brittle: they exhibit limited capacity to transfer knowledge across tasks and require massive amounts of expensive labeled data to be trained properly on a given task.

The overall aim of this project is to augment labeled data with samples of human brain activity to train DNN, which performance will generalize across a range of tasks drawn from different cognitive domains. The specific aims and hypotheses of the project are as follows:

1. Train DNNs to generate brain-like dynamics during naturalistic stimuli (i.e. movies and video games). The hypothesis is that large DNNs can be trained to accurately reproduce brain dynamics (auto-encoding), by learning connections in a space restricted by priors on anatomo-functional

connectivity generated using individual experimental measures.
2. Train DNNs to jointly generate brain-like dynamics as well as perform supervised learning tasks in vision, language and memory domains. The hypothesis is that biological brain data will help train DNNs faster on the supervised task, using more limited number of training samples than in the absence of biological brain data.
3. Transfer a DNN trained on a specific task to a task from a different domain, e.g. from vision to language. The first layers of the DNN will be identical for both task, but the highest layers will be trained in a task-specific fashion. The hypothesis is that, by using biological brain data in the training, the underlying architecture will be flexible even if initially trained in a specialized domain. Transfer learning will therefore work efficiently, achieving accurate and fast learning.

Theoretical background

Artificial deep neural networks. DNNs are mathematical models loosely inspired by biological networks. The type of DNNs that will be used in this project are trained using mathematical optimization tools to perform *supervised* tasks. This typically means that a large number of labeled data is available, e.g. the ImageNet[1] (Fei-Fei et al., 2010) database which features pictures organized by categories. The parameters of the DNNs are iteratively adjusted such that it is able to match data points, e.g. images, to labels, e.g. image categories. These artificial neural networks are called "deep" because they are structured with many layers of formal neural units, while originally only "shallow" architectures with a few layers had been successfully trained. The AlexNet[2] architecture was a turning point, demonstrating that a deep network featuring 8 layers could substantially improve the state-of-the-art on ImageNet (Krizhevsky et al., 2012). Both the number of layers and the performance have kept increasing since, with networks including tens of millions of parameters and up to 1000 layers attaining very low error rates on ImageNet (He et al., 2016).

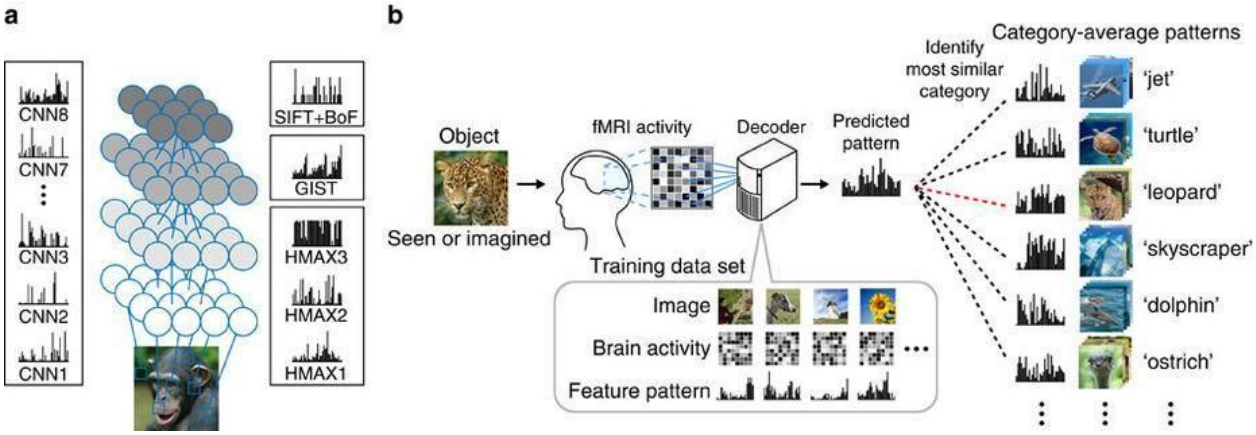

Fig. 1 - a: artificial neural network trained for image recognition on ImageNet. Each unit in each layer of the network has a specific response to each image. **b**: a regression model is used to predict the response of the artificial network from the functional activity of a human brain, exposed to the same stimuli. This predicted response can also be used to extract a predicted label from the artificial network, and implement a brain decoding model. Image from (Horikawa and Kamitani, 2017).

On the relationship between artificial and biological neural networks. Because of the striking ability of DNNs to solve tasks at a level of performance close to humans, a wave of recent works have aimed at relating the features learned by DNNs with human brain activity. The general idea is to expose a DNNs and a human participant to the same categories of stimuli, and then compare the responses of different parts of the networks across the artificial and biological networks. This type of approaches has notably been applied in vision tasks with fMRI (Horikawa and Kamitani, 2017), and magnetoencephalography in humans (Güçlü and van Gerven, 2015), as well in electrophysiology in monkeys (Yamins and DiCarlo, 2016). In particular, Horikawa and Kamitani were able to predict a substantial part of the activity of a DNNs trained on ImageNet from human fMRI data and use that prediction to infer the category of images presented to the participants beyond chance level from neuroimaging data only. But the connection between DNNs trained solely on data labels and real brain networks will necessarily remain limited (Dong et al., 2018). A natural next step would be to enforce some similarity between the response and those observed with human brain activity. This approach is promising: in a recent work, a DNN was trained for image labeling by weightening more the images that are easy to categorize based on the associated brain response, which lead to an increase in performance of the DNN (Fong et al., 2018).

Towards transferable networks. Despite impressive progress in the field of AI research, some obstacles remain. First, training DNNs to high level of accuracy still typically requires large amounts of high quality labeled data available, e.g. in biomedical applications (Ching et al., 2018). Second, a DNN trained on one task will not necessarily transfer with good performance even on a closely related task. For this reason a current area of interest is few-shot learning, i.e. the ability to learn from only a

[1] www.image-net.org/
[2] https://en.wikipedia.org/wiki/AlexNet

couple examples, facilitated by the ability to transfer knowledge gained on previous task, e.g. (Nichol et al., 2018). Interestingly, although we discussed work related to artificial and biological neural networks predominantly in the field of vision research, similar lines of work have been explored with auditory (Kell et al., 2018) and language processing (Jain and Huth, 2018). In the brain, memory and sensory representations are distributed and inter-twinned. Training of DNNs under constraint of biological data therefore appears as a promising avenue to establish architectures capable of good transfer properties, possibly across different cognitive domains.

Overview of the project. In this project, we are planning to scan extensively a handful of participants in a variety of cognitive tasks. The scanning will occur weekly in functional MRI over a year (roughly 50 scanning sessions), and bi-weekly in MEG (roughly 10 scanning sessions). The fMRI and MEG techniques were selected because of their complementary insights into brain function: good spatial resolution for fMRI, excellent temporal resolution for MEG. The core of each session will consist of watching a series of videos, composed of short clips selected to cover a variety of actions, as well as longer movies featuring more extensive narrative. These data will let us develop and validate a new class of DNNs able to capture brain dynamics efficiently, while scaling to large numbers of parameters (Aim 1). Each subject will also perform tasks drawn from, or inspired by, recent AI research. These tasks cover the memory, vision and language domains. By collecting extensive task battery on the same subjects as well as movie, we will be able to test if the DNNs trained on the movie data have extracted features that can be used to achieve good performance on a variety of tasks (Aim 2). We will also be able to test if training on one of the task will increase the speed of learning on the other tasks as well (Aim 3).

We will also include a video game task. It has been shown that skills acquired or trained through playing video games, can be transferred to various cognitive tasks relevant for everyday life. In a meta-analysis, Powers et al. (2013), found that video game play has an effect on auditory and visual processing skills, motor skills, and spatial imagery (refs). Similarly, Boot and colleagues (Boot et al. 2008) reported that action based video game players outperform non-video game players on visuospatial and working memory task, and similar results were found after action based video game training (Boot et al. 2008; Blacker et al. 2014, Green and Bavelier 2006, Green and Bavelier 2007). As such, playing video games can be considered a type of naturalistic stimuli, as it engages and strengthens various cognitive functions useful in day to day functioning. By contrast, DNNs have difficulties transferring skills acquired in one video game to another one, even when the new game shares very similar mechanics with the original game used for training (Nichol et al., 2018). By collecting extensive video game play in one participant, across various types of video games, we will be able to test if the DNNs trained on the data from one video game, have extracted features that can be used to achieve good performance on a variety of cognitive tasks, as well as on different types of video games (Aim 2). We will also be able to test if training on one of the videogame will increase the speed of learning on the other videogames, or cognitive tasks, and thus achieves superior generalization abilities compared to existing DNN architectures (Aim 3).

We want to replicate fMRI paradigm used in the Human Connectome project (HCP)[3]. HCP is a publically available data set, and their latest fMRI data release includes data acquired on 1200 participants, during 21 different tasks, spanning across 7 cognitive domains. Each task last only a few minutes, and in total the paradigm takes only 1h to scan, and most of the participants were only scanned twice on the paradigm. Both their stimuli, and Eprime[4] scripts readily downloadable, and as such a replication can be done quickly (i.e. we don't need to develop stimuli as with we are currently doing for task mentioned above). Contrary to HCP, and in line with Neuromod's aims, we will scan a small number of participants twenty times on the paradigm, giving is good dataset to start testing models for aims 2 and 3, for various cognitive domains, and also enabling us to test the reliability of activations in a particular cognitive domain.

Impact and outcomes. The main outcome of this project will be a new framework to train artificial networks to better generalize to new tasks, compared to what is currently possible.-Even if the artificial networks were to not improve on the state of the art for established supervised tasks, they may still be useful as computational models of brain networks, provided that they explain a significant portion of variance in biological data. Finally, the amount of individual longitudinal data accumulated in this project will be unprecedented, even in its first year of existence. This dataset will open new avenues to

[3] https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release
[4] https://pstnet.com/products/e-prime/

explore the longitudinal variations of functional brain networks in individual subjects, beyond the specific objectives of the study.

Methods

Subjects

As part of the main project, we will recruit about 12 participants, over the age of 18 to join the study for a minimum of five years. Participant's will be recruited by word of mouth, as we hope to find participants that will be able to commit to the extensive time commitment required by the study, and that are likely to remain throughout the course of the project. Exclusion criteria include visual or auditory problems that would prevent participants from seeing and/or hearing stimuli in the scanner, chronic or recent (i.e. in the last two years), psychiatric or neurological problems. In addition, the standard exclusion criteria for MRI and MEG will apply.

In addition, we will recruit 6 participants, to act as a control group for the *auditory battery (*see section below). These participants will be tested on the *auditory battery* at the same frequency as original 12, only they will not undergo scanning. The goal of these controls is to account for any potential changes in auditory iperformance related to learning or practice, particularly on the *MATRIX test* (see below), where the same stimuli will be repeated at each session.

Auditory sessions

There will be two types of auditory sessions; *auditory battery*, and the *auditory attenuation pilot*.

The **auditory battery** is designed to ensure that we can monitor, and quantify any temporary, and/or permanent change in participants' auditory systems as a result of repeated exposure to different types of scanning related noise. The auditory battery will be done approximately once a month, or every four scans, and will last around 60 minutes. All hearing tests will be carried out in a soundproof booth and will be administered by trained staff. All test are non-invasive, and carry no risk to the participants.

Hearing tests included in our battery:

Pure-tone Audiometry (20 min): This test measures hearing acuity for tones at low to high frequencies (Hz), and at different intensities (dB), (Huizing, 1951). Participants will be asked to wear a set of headphone and instructed to raise their hand when a sound is detected. For every correct answer the volume will be lowered, and for every incorrect answer the volume is increased. This process is repeated until a participant responds to a test signal (or tone), 70% of the time on ascending trials (i.e when we establish their threshold). Thresholds will be measured for a range of both low, and high frequencies (500-1600 Hz).

Tympanometry (5 min): This tests the state of the middle ear, namely the mobility of the eardrum (tympanic membrane), and the conduction bones. Following a visual inspection of the ear to ensure the path to the eardrum is clear, a tympanometer probe will be inserted into the participants ear canal. The probe will then alters the pressure in the ear canal, send out a pure tone, and measure the eardrum mobility to the sound at different pressures. Tympanometry can be used to test the condition of the middle ear, determining whether it is functioning normally, contains fluid, or whether the tympanic membrane has been punctured (Lilly, 1984). This test will also allow us to differentiate between temporary hearing loss due to sinus infections or allergies (Lazo-Sáenz et al., 2005), versus potential loss due to scanning conditions.

Otoacoustic emissions - (OAE's; 15 min): OAE's are sounds given off by hair cells in the inner ear that respond to sound by vibrating (Kemp, 2002). The vibration produces a very quiet sound that echoes back into the middle ear. If you have normal hearing, you will produce OAEs, however if your hearing loss is greater than 25–30 decibels (dB), you will not. OAE's are measured by placing a small earphone/probe in the outer cavity of - your ear, the probe then output sounds into your ear and measures the sounds that come back.

English and French MATRIX speech test (10 min): MATRIX speech test will assess a participant's ability to understand sentences while exposed to background noise (Hagerman, 1982). Participants will be asked to wear a set of headphone that will deliver one of 20 different sentences with background

noise. Participants will be asked to repeat the sentence they just heard. If they only partially heard the sentence, they will be asked to repeat the portion of the sentence or words they heard. Participants will be tested on the French or English test based on which language they are more proficient in.

The **sound attenuation pilot** (60 minutes), will be used to quantify the efficiency of MRI compatible ear muffs we built at UNF. Each participant will have their free-field thresholds tested in the following conditions: 1) with nothing, to establish their baseline, 2) with the sensimetrics earbuds (MR-compatible sound delivery system available at UNF that provides about 20 dB of sound protection at higher frequency[5]), 3) with earbuds and headcase (see Personalized headcase section), 4) with earbuds, headcase and earmuffs. During the assessment, the participants will sit in a soundproof booth, facing two speakers, located on either side of their chair, at an approximately 45 degree angle. Sounds will be presented to the participants via the two speakers, and participants will be instructed to press a button every time they hear a sound. Similarly to the Pure-tone audiometry (see above), free-field thresholds measure hearing acuity for tones at low to high frequencies (Hz), and at different intensities (dB), only sounds are presented via the speakers instead of headphones.

Personalized headcases

In order to minimize movement during neuroimaging scans, for each participant we will purchase a custom-designed, personalized headcases from Caseforge[6]. In order to personalize the headcases, we will need to scan each participant's head using a handheld 3D scanner. The images will be sent to Caseforge's secure servers where they will be used to mill two personalized headcases for use in the MRI, and MEG scanners.

The **motion task** (60 minutes), the goal of this task is to quantify the amount of head motion motion with and without the personalized headcase. Small Qr codes, called ArUco[7], will be places on the participants' head, and motion will be recorded using an MRI compatible camera. Participants will be scanned in both the fmri and MEG, with and without their headcases, doing a variety of tasks. These tasks will include watching resting state, reading a text out loud, watching a video, and playing a video game.

Autonomic measures

Participant's pulse will be measured using a MR-compatible plethysmograph, placed on the ankle to obtain beat-by-beat estimates of heart rate based on the interval between successive systolic peaks. Skin conductance will be measured using two MR-compatible electrodes applied to the sole of the foot. An electrocardiogram will be recorded using three MR-compatible electrodes places on the participants upper body and will measure the electrical activity generated by the heart. A pneumatic belt will be place under the participants ribs, to measure their rate of respiration. Measures will be acquired at 1000 Hz using the BIOPAC[8] systems available at l'UNF, and data will be processed in an event-related manner. The changes in heart rate induced in the 10 seconds following the onset of the stimuli will be measured relative to baseline (1 sec pre-stimulus). The skin conductance response will be assessed using the area under the curve (10 sec epoch post-stimulus) of the smoothed (1-sec average) and drift-corrected (1-sec differential) signal. These measures will be collected in during both the fMRI and MEG scans.

Visual presentation, eye tracking and pupil dilation

During fMRI scans, all visual stimuli will be presented in a standardized fashion: where they will see stimuli presented screen at the the back of the scanner via a mirror. During the MEG scan stimuli will be presented on a screen[9], and eye motion and pupil dilation will be recorded using using and MEG compatible Eyelink.

Pupillometry and eye tracking data will and analyzed using the image-analysis software ViewPoint from

[5] http://www.sens.com/products/model-s15/#documentation
[6] https://caseforge.co/
[7] https://docs.opencv.org/3.1.0/d5/dae/tutorial_aruco_detection.html
[8] https://www.biopac.com/application/magnetic-resonance-imaging-with-biopac-equipment/
[9] https://www.sr-research.com/solutions/fmri-meg-solutions/

Arrington Systems [10], or an equivalent open source software.

Facial expression stimuli

During both the fMRI and MEG scans participants facial expressions will be videotaped using MRI/MEG compatible cameras, that will either be mounted on the head coil, or on the wall of the MEG room. The facial responses will be quantified by trained personnel using the Facial Action Coding System (Ekman, 1980).

Experimental Procedures

Participants will attend a recruitment session where they be given the consent form to read over. The exclusion criteria, and MRI & MEG screening criteria will also be checked. Upon acceptance of participants, the screening questionnaire, will be given to the subject, and a three-dimensional image of their head will be acquired using a Caseforge hand-held scanner. Future scanning and testing session will be scheduled for the upcoming month.

Structural MRI sessions (60 minutes) will be acquired approximately once every 3 months. The purpose of these session is to collect various types of structural information. Using the 64-channel head coil and standard UNF sequences, we will acquired the following types of scans: T1 ME-MPRAGE (9 min), Diffusion-weighted scan (6 min), MTstat (6 min), MP2RAGE (5 min), T2 -SPACE (6 min), PD-weighted images (10 min). During the scans participants will be instructed to relax, and try to remain still.

fMRI sessions (60 minutes). Images will be acquired on the 64-channel head coil using a MB sequence developed at l'UNF. Participants will be asked to participate in two type of tasks that will last about 30 minutes; during the video task they will watch videos of vary lengths, and the video game task (see below). The other task will be one of the following:

The **language task** (13 minutes). Participants will be presented word triplets; one target word, followed by two words shown side by side. Participants will be instructed to selected which of the two words are most similar to the target word. They will use an MR-compatible mouse to record their choice (a - left words; - right word). Stimuli will be presented using a randomized inter-stimulus interval, and during the interstimulus interval participants will be presented a blank screen.

The **reading task** (15 minutes). During this task, 10 participants will be asked to read out loud a text that they see on the screen. This task will be used to quantify motion and to assess the efficacy of the headcase at reducing motion.

The **image task** (13 minutes). Participants will be presented with images depicting real-world scenes, and judge whether they liked, disliked, or were neutral about the image. An MR-compatible mouse will record their choice. Images are selected from a large scale and diversity image datasets such as ImageNet[11], Common Objects in Context dataset[12], SUN database (Xiao et al., 2010), and the THINGS dataset. We selected images to cover a wide range of categories, based on ImageNet, with at least two different examples in each category. Images will be presented in event-related fMRI
The **memory task** (13 minutes), participants will be presented a randomized sequence of everyday images (selected from ImageNet), and control stimuli (an abstract multicolored image) for three seconds each. Images will appear in one of the four quadrants of the screen and participants will be asked to remember the images, as well as their position in the four quadrants (either top-left, top right, bottom-left or bottom-right). To control for attention, they will also be asked to press a key each time they see a stimuli. The encoding phase will last 13 minutes (Belleville et al., 2014). The retrieval will be done outside the scanner, and 10 minutes after the end of the encoding phase. Participants will be presented the same set of images from the scanner, in a new order, along with new images, all appearing in the center of the screen. They will have to determine whether each image has been previously presented, and if YES, where it was located (Belleville et al., 2014).

During **video game task** (30 to 60 minutes depending on the session type) participants will use an

[10] http://www.arringtonresearch.com/index.html
[11] http://www.image-net.org/
[12] http://cocodataset.org/#home

MRI compatible video game controller built by André Cyr, the UNF engineer. The video game controller uses fiber optics to minimize the likelihood of artifacts in the MRI scanner, and the shell of the controller is 3D printed using an in-house printer. While being scanned participants will be to play video games, such as Shinobi III[13]. The games, as well marker of the participants' style of play (i.e. button presses, number of jumps, runs, lunges, ducks, etc), and game stats will be recorded and saved for later analysis.

Resting state scan (14 minutes; two seven minute runs), will be acquire approximately once every four sessions, and participants will be presented with the Inscapes[14] video to help reduce movement, minimize cognitive loads (Vanderwal et al., 2015).

MEG sessions (120 minutes), will mimic fMRI sessions. In the first task they will watch 30 minutes of videos of various lengths and during the second portion they will be assigned one of the following tasks: *language task*, *images task*, *memory task, or video game task*. Participants will also do the resting state. The exact nature of the tasks are described above.

Pilot Project (60 minutes of 11 visits**).** Four participants will undergo fmri 10 session of 1h and 1 MEG session during which they will do the HCP fMRI protocol[15]. This task will only be acquired in the MRI scanner. The protocol consists of seven tasks, namely; gambling, motor, language processing, social cognition, relational processing, emotional processing, and working memory. The descriptions below were adapted from the HCP protocol site. Before each task participants are given detailed instructions, and are given examples, as well as a practice run.

During the **gambling task** (approximately 4 minutes), participants will play a card guessing game where they will be told to guess whether the number of a card (represented by a "?" presented for 1500*ms* on the screen) is above or below 5 (Delgado et al. 2000). They will indicate their choice using button press and following their choice they will be shown the correct number. If they guess correctly they win money ($1.00 - reward condition), if they guess incorrectly they lose money ($0.50 - loss condition), and if the number is exactly 5 they will neither win or lose money (neutral condition). The conditions are presented in blocks of 8 trials that are either mostly reward (6 reward trials pseudo randomly interleaved with either 1 neutral and 1 loss trial, 2 neutral trials, or 2 loss trials) or mostly loss (6 loss trials pseudo-randomly interleaved with either 1 neutral and 1 reward trial, 2 neutral trials, or 2 reward trials). There are four block per run (2 mostly win and 2 mostly loss), and two runs in total. At the end of the session, 5 trial will be randomly selected and participants will corresponding amount of money that they won/loss on those 5 trials.

During the **motor task** (approximately 5 minutes), adapted from (Buckner et al. 2011; Yeo et al. 2011), participants will be presented a visual cue, and asked to either tap their left or right fingers, squeeze their left or right toes, or move their tongue to map motor area. Each movement lasts 12 seconds, and in total there are 13 blocks, with 2 of tongue movements, 4 of hand movements (2 right and 2 left), and 4 of foot movements (2 right and 2 left), and three 15 second fixation blocks where they will be instructed not to move anything. There are two runs in total, and 13 blocks per run.
During the **language processing task** (approximately 5 minutes), participants either listen to an auditory story (5-9 sentences, about 20 seconds), followed by a two-alternative forced-choice question, or they listen to a math problem (addition and subtraction only, varies in length), and instructed to push a button select the first or the second answer as being correct. The task is adaptive so that for every correct answer the level of difficulty increases. The math task is designed this way to maintain the same level of difficulty between participants. There are 2 runs, each with 4 story and 4 math blocks, interleaved.

During the **social cognition task** (approximately 5 minutes), participants will be presented with short video clips (20 seconds) of objects (squares, circles, triangles) that either interact in some way, or move randomly on the screen (Castelli *et al.* 2000; (Wheatley *et al.* 2007). Following each clip participants will be asked to judge whether the objects had a "Mental interaction" (an interaction that appears as if the shapes are taking into account each other's feelings and thoughts), whether the are "Not Sure", or if there was "No interaction". Button presses are used to record their responses. In each of the two runs, participants will view 5 "Mental" videos and 5 random videos and have 5 fixation

[13] https://en.wikipedia.org/wiki/Shinobi_III:_Return_of_the_Ninja_Master
[14] https://vimeo.com/67962604
[15] http://protocols.humanconnectome.org/HCP/3T/task-fMRI-protocol-details.html

blocks of 15 seconds each.

During the **relational processing task** (approximately 5 minutes) participants will be shown 6 different shapes filled with 1 of 6 different textures (Smith *et al.* 2007). The are two conditions: *relations processing,* and *control matching condition.* In the *relational processing* condition, 2 pairs of objects are presented on the screen, with one pair at the top of the screen, and the other pair at the bottom. They are instructed to decide what dimension differs in the top pair (shape or texture), and then decide if the bottom pair differ, or not, on the same dimension (i.e. if the top pair differs in shape, does the bottom pair also differ in shape). Their answers are recorded by one of two button presses: "a" differ on same dimension; "b" don't differ on same dimension. In the *control matching* condition, participants will be shown two objects at the top of the screen, and one object at the bottom of the screen, with a word in the middle of the screen (either "shape" or "texture").They will be told to decide whether the bottom object matches either of the top two objects on that dimension (i.e., if the word is "shape", is the bottom object the same shape as either of the top two objects). Participants respond "yes" or "no" using the button box. For the relational condition, the stimuli are presented for 3500 ms, with a 500 ms ITI, and there are four trials per block. In the matching condition, stimuli are presented for 2800 ms, with a 400 ms ITI, and there are 5 trials per block. In total there are two runs, each with three relational blocks, three matching blocks and three 16-second fixation blocks

During the **emotion processing task** (approximately 4 minutes) participants will be shown triads of faces or shapes, and asked to decide which of the shapes at the bottom of the screen matches the target face/ shape at the top of the screen (adapted from Smith *et al.* 2007). Faces have have either an angry or fearful expression. Faces, and shapes are presented in three blocks of 6 trials (3 faces and 3 shapes), with each trial lasting 2 seconds, followed by a 1 second interstimulus interval. Each block is preceded by a 3000 ms task cue ("shape" or "face"), so that each block is 21 seconds including the cue. In total there are two runs, three face blocks and three shape blocks, with 8 seconds of fixation at the end of each run.

During the **working memory task** (approximately 5 minutes) there are two subtasks in the paradigm; a category specific representation, and a working memory task. Participants are presented with blocks of either places, tools, faces, and body parts. Within each run, all 4 types of stimuli are presented in block, with each block being labelled as a 2-back task (participants need to indicate if they saw the same image two images back), or a version of a 0-back task (participants are shown a target at the start of the trial and they need to indicate if the image they are seeing matched the target). Each image is presented for 2 seconds, followed by a 500 ms ITI. Stimuli are presented for 2 seconds, following by a 500 ms inter-task interval. Each of the 2 runs includes 10 trials, and 4 fixations blocks (15 secs).

Resting state. In every other session, one 15 minutes rfMRI run will be acquired. Participants will have their eye open, be looking at fixation cross in the middle of the screen and be instructed to not fall asleep.

Brain Imaging preprocessing

All fMRI and MEG runs will undergo standardized preprocessing pipelines.

Briefly, fMRI volumes will be realigned using rigid body transform, within and between functional runs (including between sessions) and corrected for differences in slice acquisition times using temporal interpolations. Functional MRI volumes will be corrected from distortions due to field inhomogeneity based on a reference field map image. Temporal confounding effects (slow time drifts, motion parameters) will be regressed from the data. Independent component analysis will be used to identify and remove effects of physiological noise, such as cardiac, respiratory and motion related fluctuations. All available T1 weighted structural scans for each subject will be registered together using a rigid body transformation, and averaged to create an unbiased, high quality reference scan. Cortical surface will be extracted from this average structural image, which will also be coregistered with the functional scan for each subject. Functional runs will finally be interpolated on the cortical surface. This preprocessing pipeline is implemented in the package called fmriprep[16], developed by members and collaborators of the Poldrack laboratory at Stanford university, and depends on a mixture of different analytical packages including Freesurfer (Dale et al., 1999), AFNI (Cox, 1996) and nilearn (Abraham et

[16] https://github.com/poldracklab/fmriprep

al., 2014). A detailed description of the fmriprep pipeline is available online[17].

Standard MEG data cleaning (ICA-based ocular, cardiac and muscle artefact rejection, and exclusion of bad segments/channels) will be followed by data segmentation with respect to events of interest. A MEG-compatible eye-tracker, in combination with EOG traces, will be used to control for unwanted artefacts or synchronizations generated by eye movements (e.g. saccades). MEG data-preprocessing and analyses will be conducted using a combination of open-source tools developed in the Jerbi lab (NeuroPycon[18], Visbrain[19], Brainpipe[20], and TensorPac[21], as well as well-established freely-available toolboxes, i.e. MNE-python (Gramfort et al., 2013).

Artificial neuronal models

Regarding aim 1, we will develop "vector-quantized variational auto-encoders" (VQ-VAE) that compress the brain images over a time window of 15 seconds, using discrete latent variables (van den Oord et al., 2017). The training of the VQ-VAE will be implemented using all the video and video game fMRI data available for each subject, and a separate model will be trained for each subject. Compressed representation of the video and audio signals, as well as text transcripts of the video will be generated using established techniques (Richard, H, Pinho, AL, Thirion, B, Charpiat, G, 2018). We will then train a sparse attentive backtracking (SAB) recurrent neural network on these discrete latent variable representation to predict future fMRI time series from past time points, using a mixture of past video, audio and functional signals. The performance of the recurrent DNN will be evaluated by the accuracy of prediction of future fMRI time points from past fMRI time points and compared to simple auto-regressive models as a benchmark, using split half cross-validation on available data (half of time points for training, half of time points to evaluate the accuracy of the model).

Regarding aim 2, the same VQ-VAE model trained for aim 1 will be trained to reduce the dimensionality of fMRI time series specifically on the data collected in the language, image, and memory tasks. A separate SAB recurrent DNN will be trained to solve the same task as required for humans. We will evaluate the ability to solve those tasks in isolation, using ten-fold cross validation to evaluate the performance of the models.

Regarding aim 3, we will use the VQ-VAE model trained on all of the movie data, and only train the SAB component of the recurrent DNN to solve each particular task. We will evaluate the performance of this variant of the model using ten-fold cross-validation (within each task) and expect this variant of the recurrent DNN to be easier to train and outperform the DNNs trained specifically on each task, as described in aim 2. This analysis will test directly that the representation learned on movies, or video games transfer well to other cognitive context (language, memory, image).

Sample size

The primary aim of this project is to demonstrate our ability to efficiently train artificial DNNs to solve a variety of task. Potentially this could be achieved with data collected on a single human participant. The limiting factor is here the amount of data available on each individual. The amount of individual data will exceed what was collected in the only comparable study we are aware of (Fong et al., 2018). We do not think we can realistically collect more data than currently planned, because of the large time demands already required from the research participants. To check that our observation generalize we would like to have three participants for each task, which we increased to four to anticipate participants potentially dropping out of the study. Because we will be testing three tasks in parallel, the core cohort will therefore include 12 subjects.

[17] https://fmriprep.readthedocs.io/en/stable/citing.html

[18] https://neuropycon.github.io/neuropycon_doc/

[19] http://visbrain.org/

[20] https://etiennecmb.github.io/brainpipe/

[21] https://etiennecmb.github.io/tensorpac/

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. 8, 14.

Belleville, S., Fouquet, C., Duchesne, S., Louis Collins, D., Hudon, C., the CIMA-Q group: Consortium for the Early Identification of Alzheimer's disease-Quebec, 2014. Detecting Early Preclinical Alzheimer's Disease via Cognition, Neuropsychiatry, and Neuroimaging: Qualitative Review and Recommendations for Testing. J. Alzheimers. Dis. 42, S375–S382.

Boot, W.,R., Kramer, A., F., Simons, D., J., Fabiani, M., & Gratton G., 2008. The effects of video game playing on attention, memory, and executive control. Acta Psychologica, 129, 387–398.

Blacker, K., J., Curby, K., M., Klobusicky, E., & Chein, J., M., 2014. Effects of action video game training on visual working memory. Journal of Experimental Psychology: Human Perception and Performance, 40, 1992–2004.

Buckner, R., L., Krienen, F., M., Castellanos, A., Diaz, J.,C., Yeo, B.,T., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J Neurophysiol*. 106(5), 2322-45.

Castelli, F., Happé, U., Frith, C., 2000. Frith Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns
Neuroimage, 12, 314-325.

Chang, N., Pyles, J.A., Gupta, A., Tarr, M.J., Aminoff, E.M., 09/2018. BOLD5000: A public fMRI dataset of 5000 images. eprint arXiv:1809.01281.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., Xie, W., Rosen, G.L., Lengerich, B.J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A.E., Shrikumar, A., Xu, J., Cofer, E.M., Lavender, C.A., Turaga, S.C., Alexandari, A.M., Lu, Z., Harris, D.J., DeCaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L.K., Segler, M.H.S., Boca, S.M., Swamidass, S.J., Huang, A., Gitter, A., Greene, C.S., 2018. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface 15.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9, 179–194.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., Fiez, J.A., 2000.Tracking the hemodynamic responses to reward and punishment in the striatum. J. Neurophysiol. 84, 3072 – 3077.

Dong, Q., Wang, H., Hu, Z., 2018. Commentary: Using goal-driven deep learning models to understand sensory cortex. Front. Comput. Neurosci. 12, 4.

Ekman, P., 1980. Asymmetry in facial expression. Science 209, 833–834.

Fei-Fei, L., Deng, J., Li, K., 2010. ImageNet: Constructing a large-scale image database. J. Vis. 9, 1037–1037.

Fong, R.C., Scheirer, W.J., Cox, D.D., 2018. Using human brain activity to guide machine learning. Sci. Rep. 8, 5397.

Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013. MEG and EEG data analysis with MNE-Python. Front. Neurosci. 7, 267.

Güçlü, U., van Gerven, M.A.J., 2015. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J. Neurosci. 35, 10005–10014.

Green, C., S., & Bavelier, D., 2006. Enumeration versus multiple object tracking: The case of action video game players. Cognition, 101, 217–245.

Green, C., S., & Bavelier, D., 2007. Action-video-game experience alters the spatial resolution of vision. Psychological Science, 18, 88–94.

Hagerman, B., 1982. Sentences for testing speech intelligibility in noise. Scand. Audiol. 11, 79–87.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. openaccess.thecvf.com, pp. 770–778.

Horikawa, T., Kamitani, Y., 2017. Generic decoding of seen and imagined objects using hierarchical visual features. Nat. Commun. 8, 15037.

Huizing, H.C., 1951. Pure tone audiometry. Acta Otolaryngol. 40, 51–61.

Jain, S., Huth, A., 2018. Incorporating Context into Language Encoding Models for fMRI. bioRxiv.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., McDermott, J.H., 2018. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98, 630–644.e16.

Kemp, D.T., 2002. Otoacoustic emissions, their origin in cochlear function, and use. Br. Med. Bull. 63, 223–241.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1097–1105.
Lazo-Sáenz, J.G., Galván-Aguilera, A.A., Martínez-Ordaz, V.A., Velasco-Rodríguez, V.M., Nieves-Rentería, A., Rincón-CastañEda, C., 2005. Eustachian Tube Dysfunction in Allergic Rhinitis. Otolaryngol. Head Neck Surg. 132, 626–629.
Lilly, D.J., 1984. Multiple frequency, multiple component tympanometry: new approaches to an old diagnostic problem. Ear Hear. 5, 300–308.
Nichol, A., Pfau, V., Hesse, C., Klimov, O., Schulman, J., 2018. Gotta Learn Fast: A New Benchmark for Generalization in RL. arXiv [cs.LG].
Richard, H, Pinho, AL, Thirion, B, Charpiat, G, 2018. Optimizing deep video representation to match brain activity. arXiv: 1809.02440.
Smith, R., Keramatian, K., Christoff, K., 2007, Localizing the rostrolateral prefrontal cortex at the individual level. Neuroimage. 36, 1387–1396.
van den Oord, A., Vinyals, O., Kavukcuoglu, K., 2017. Neural Discrete Representation Learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 6306–6315.
Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., Xavier Castellanos, F., 2015. Inscapes : A movie paradigm to improve compliance in functional magnetic resonance imaging. Neuroimage 122, 222–232.
Wheatley, T., Milleville, S., C., Martin, A., 2007. Understanding animate agents: distinct roles for the social network and mirror system. Psychological Science. 18, 469–474.
Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A., 2010. SUN database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
Yeo, B., T., Krienen, F. ,M., Sepulcre, J., Sabuncu, M., R., Lashkari, D., Hollinshead, M., Roffman, J., L., Smoller, J., W., Zollei, L., Polimeni, J.,R, Fischl, B., Liu, H., Buckner, R., L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. Journal of Neurophysiology. 106, 1125–1165.
Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365.