Computational Cognitive Neuroscience: Assignment 3 Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness

Xiao Heng (s2032451)

February 22, 2024

Background

This essay is mainly based on the research of Cullen et al. (2018), which shows that active inference enables simulated subjects to actively develop detailed representations of gaming environments through the epistemic and value-based targets. it also demonstrate the use of a principled algorithmic and neurobiological framework for testing hypotheses in psychiatric illness.

Reinforcement-learning (RL) models with model-based functional magnetic resonance imaging (O'doherty et al., 2007) have predominated in this area, with neuromodulators and decision-making circuits in the striatum (Daw et al., 2002) highlighted as crucial neural substrates relating aberrant decision making and learning to psychiatric illness symptoms. Meanwhile, hierarchal Bayesian analysis has also been applied in conjunction with model-based functional magnetic resonance imaging to model prediction errors (Iglesias et al., 2013), focusing on errors related to prior beliefs in terms of general real-world states, which is one key distinction compared with the previous RL models whose concern is rather prediction errors on reward or punishment.

In terms of computational psychiatric approaches, normative models and process models of the brain are commonly considered. The former one focus on the computational or mathematical target of the brain, regardless of implementing process; while the latter one concerns more about the implementing mechanism of a specific algorithm. However, active inference comprises both. It minimizes the free energy (which is a potential internal target of the brain) while also proposing distinct neurobiological components to implement the optimization process via distinct message-passing sequences between prefrontal, sensory, and neuromodulatory neurons (Friston et al., 2017), usually with the help of variational Bayesian inference, including some more detailed skills such as Kullback-Leibler divergence. In fact, the method of active inference has been applied within a large range of problems in the field of cognitive neuroscience and neuropsychiatry. As for the free energy principle, it is based on a Bayesian idea that the brain could be viewed as an inference engine. Under such proposition, any behaviour from a given system will be modelled, and the target is to minimize the difference between the models of the world and the sense with associated perceptions. The term 'surprise' could be used to describe this difference, which will be updated and minimized from the continuous observations and corrections during the interactions with environments. In a more formal definition, free energy could be described an information theory measure that bounds or limits (by being greater than) the surprise on sampling some data, given a generative model (Friston, 2010). Note that, the free energy principle could really

expain the expression that 'life feeds on negative entropy', stated by Schrödinger, to some extent.

Methods

The paper focus on the way to test and explain mechanistic and algorithmic properties of psychiatric disorders by applying active inference models on game platform, which contains two main topics (exemplar analyses):

- aging on cognition;
- putative feature of anhedonia (diminished sensitivity to reward).

Now, begin with a brief introduction of the simulation environment, the *Doom* environment on OpenAI Gym. *Doom* is a pseudo–three-dimensional game. The target ('end goal') of the game is 'shooting the monster'. The action space is {'move right', 'move left', 'fire'}. Considering the relative location between agent and monster, the 6-state space is then {'left-of-monster, not firing', 'left-of-monster, firing', 'right-of-monster, not firing', 'right-of-monster, firing', 'centered-on-monster, not firing', 'centered-on-monster, firing'} (there is also 10-state space later, with additional {'far-left-of-monster, firing', 'far-left-of-monster, not firing', 'far-right-of-monster, firing', 'far-right-of-monster, not firing'}). Note that, the 'state' here could be regarded as 'state-action' pairs (although the actions are only classified as 'firing' or 'not firing'). Such definition makes it more convenient for constructing state transition matrix later.

In the experiment, three variables are constructed for learning: matrix A, matrix B and vector C. Matrix A comprises the agent's belief of mapping between sensory information and states of environment. To simplify, Harris corner-detection algorithm (Harris et al., 1988) is applied to deconvolve the pixel data into a set of manageable corner features. The Harris corner-detection algorithm captures corner feature in images by calculating derivatives and comparing sum of squared differences between image patches, which is commonly used in computer vision field. Then matrix B comprises the agent's beliefs of state trasitions under the whole action space (as described previously). Finally, vector C comprises beliefs about expected outcomes, which is a proxy for utility or reward. In the learning scheme, specifically, matrix A is set to the identity, assuming that there is no difference for agents in recognizing the image visually. Matrix B, representing trasition probabilities, is unknown, and is initialized as 'flat' prior under Dirichlet distributions, which is commonly used as prior distribution in Bayesian-related statistics, indicating no de-novo knowledge of agents about the environment. As for vector C, it is set as C=$[0.1, 0, 0.2, 0.6, 0.1, 0]$ under 6-state condition and C=$[0.05, 0, 0.05, 0, 0.2, 0.6, 0.05, 0, 0.05, 0]$ under 10-state condition respectively, implying a simple prior that the agent should 'fire' when 'centered-on-monster' with high probability; and when the relative location is not at middle, 'not firing' should be more preferable.

The feature extraction methods are illustrated in **figure 1**.

For each artificial agent, the simulation (learning process) is conducted for 128 episodes, and the evaluation is from the average performance of 50 runs. Here, the 'episode' is defined as the period from initialization to the end of game, triggered by achieving the target (shooting the monster) or 10 seconds (350 frames) of game time elapsing. For the reward, if action is 'fire' and the relative location is 'centered-on-monster', 100 points of reward will be returned, and then game is over. Before the end, each time step will lead to -1 point (penalty), or an incorrect 'fire'.

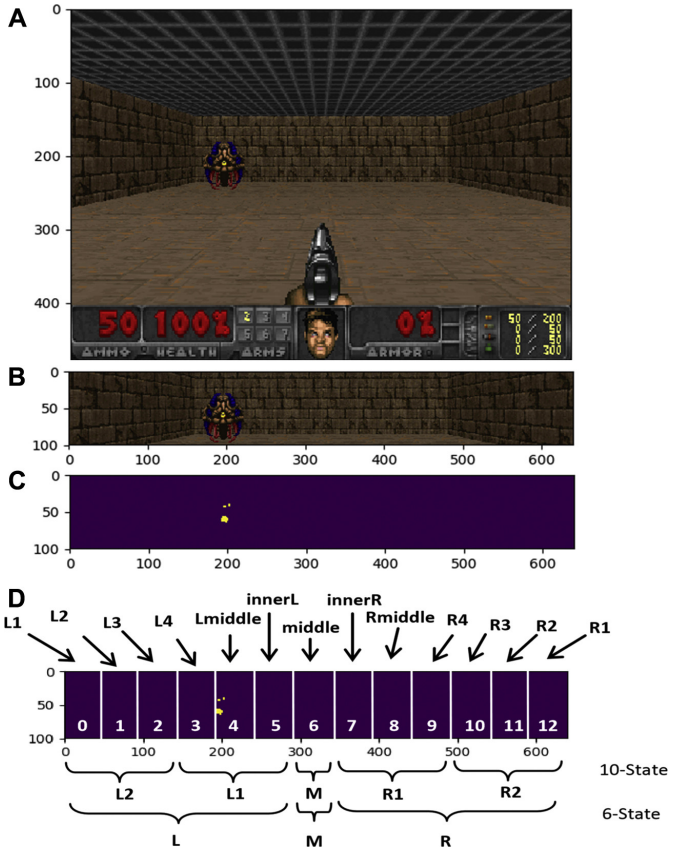

Figure 1: Feature extraction. **(A)** Observation from *Doom* in 480×640 pixel space. **(B)** By removing redundant parts, the 100×640 pixel space is adopted. **(C)** Output from the Harris corner-detection algorithm with highlighted local maxima of the corner response function. **(D)** discrete states by the location (in pixels) of the target (monster) relative to the player and whether the agent is currently shooting.

To simulate the neuronal responses and examine the putative neurobiological correlates of free energy minimization during episodes, the steps are as follows:

- before each action being selected, the agent will estimate current and future states based on experience and action-dependent state transitions;
- during the state estimation, iterative gradients (assuming gradient descent) are applied for optimization;
- record all the gradient based estimates, assuming the fluctuations among them are directly mappable to firing in the prefrontal cortex;
- to simulate an putative local field potential (LFP), simply plot the estimates after a band-pass filtering process with a working time update of 16 msec;
- to generate an associated blood oxygen level–dependent (BOLD) response, pass the simulated LFP through another function of neurovascular coupling that maps local electrical

brain activity to changes observed using fMRI, which is based on a hemodynamic response function (Buxton et al., 1998).

Besides machine learning in simulation, the research also collated scores from real human players. As for the filtering during recruitment, the screening of subjects depends on self-report of no psychiatric or neurological history via a participant information sheet. Finally, 16 players with equal gender proportion and 37 ± 17 (mean $\pm$ SD) years of age are selected, whose performance would be measured. Specifically, in the research of 'aging on cognition', the full sample should be divided into two groups with significantly different ages: the younger group ($n = 9$, 22 ± 1 years, matches the 10-state agents) and the older group ($n = 7$, 56 ± 5 years, matches the 6-state agents). The rationale behind this comparison of 6-state and 10-state agents representing people of different ages is based on the research of free energy principle and aging by Gilbert et al. (2016). In this case, there are enough materials that support the proposition that synaptic loss over life span may lead to adaptive pruning. This will make brains of the older more resilient to short-term changes with environmental input.

For each participant, 64 consecutive episodes are conducted and measured. Note that, three buttons (action space) are {'left', 'right', 'space'} while they are not mapping deterministicly into the original function of {'move-left', 'move-right', 'fire'}. However, the mapping are set randomly for each participant so that they need to learn the correct button associated with certain action (since the simulated agents also do not know the different of action 0, 1, 2, the 'flat prior' on the actions is secured).

In terms of the two different optimization targets during learning (training), first we look at the policy updating under active inference. Due to the fact that the original paper does not give the deduction process, we cited and supplemented the lacked detail from Friston et al. (2017) as follows:

$$\begin{aligned} G(\pi, t) &= E_{\tilde{Q}}[\ln Q(s_t|\pi) - \ln P(s_t, o_t|\tilde{o}, \pi)] \\ &= E_{\tilde{Q}}[\ln Q(s_t|\pi) - \ln P(s_t|o_t, \tilde{o}, \pi) = \ln P(o_t)] \\ &\approx E_{\tilde{Q}}[\ln Q(o_t|\pi) - \ln Q(o_t|s_t, \pi)] - E_{\tilde{Q}}[\ln P(o_t)], \end{aligned}$$

in which: π is the policy; s is the state; o is the outcome; $\tilde{o} = (o_1, \ldots, o_t)$ represents all the past experience and information from time step t; the $G(\pi, t)$ is the expected free energy of policy π associated with time step t; $\tilde{Q} = Q(o_t, s_t|\pi) = P(o_t|s_t)Q(s_t|\pi) \approx P(o_t, s_t|\tilde{o}, \pi)$; $Q(o_t|s_t, \pi) = P(o_t|s_t)$.

In the last line of the equations, the first expectation term is negative 'epistemic value', and the second expectation term (regardless of the minus symbol) is 'extrinsic value', and our optimization task is to minimize the $G(\pi, t)$.

However, back to the discussed paper, they use a slightly different expression:

$$Q(\pi|t) =< \ln P(o|s) - \ln P(o|\pi) > + < \ln P(o) >$$

They define $Q(\pi, t)$ as the negative expected free energy (so that the optimization turns to be maximization, and so that all variables are multipled by -1); and they consider the policy π at given time step t (conditional on t), so that all the subscripts t are removed. Now in this expression, still, the first term is (positive) 'epistemic value', and the second term is 'extrinsic value'.

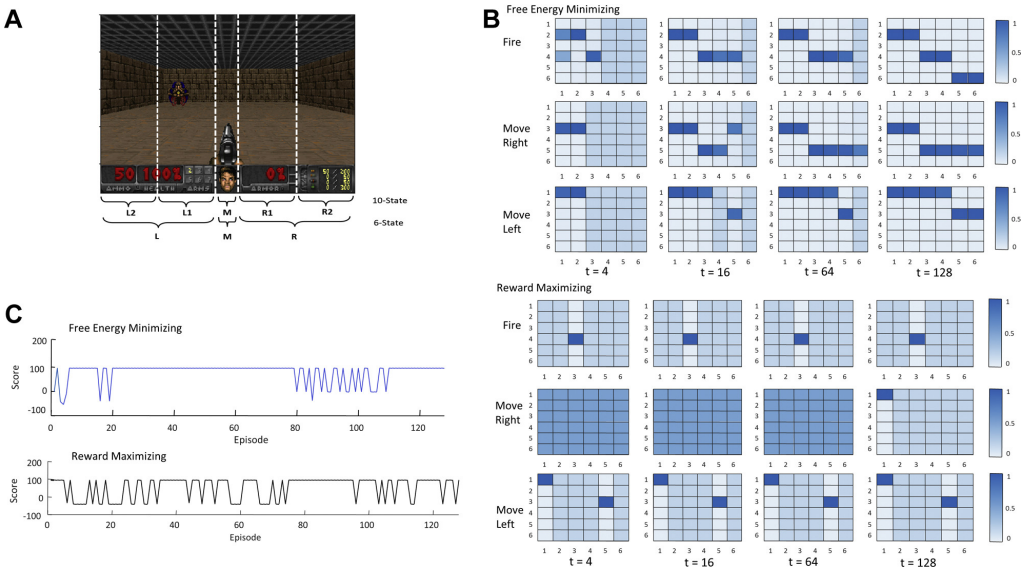

Figure 2: Adaptive behaviors and learned contingencies.

Now, given this expression, when it comes to the construction of reward maximising target, the only thing we need to do is simply to remove the 'epistemic value' term from the evaluation of the policy as above.

Results

To analyze the two topics above, four comparisons are conducted:

- Free energy–minimizing vs. reward (goal)–maximizing;
- Active inference simulation vs. humans;
- Complex model (10-states, younger) vs. simple model (6-states, older);
- Anhedonic priors vs. motivated priors.

Free energy–minimizing vs. reward (goal)–maximizing

Before futher analysis, firstly, the choice of target optimization method should be compared. In this paper, two main methods are considered: the free energy–minimizing by active inference which has nice biological explanations, and the reward–maximizing which is the most classic and basic idea in RL algorithms. The comparison results are shown in **figure 2**.

Here only the two different optimization targets are considered, so all other settings are kept the same under the 6-state space. In **figure 2B**, it describes matrices B of the two agents at time steps 4, 16, 64 and 128 respectively. This shows a visualization of emerging learning process of state-action dynamics through trials within the environment. Note that, the learned state transitions under reward maximization are clearly less robust and underfitting, indicated by the

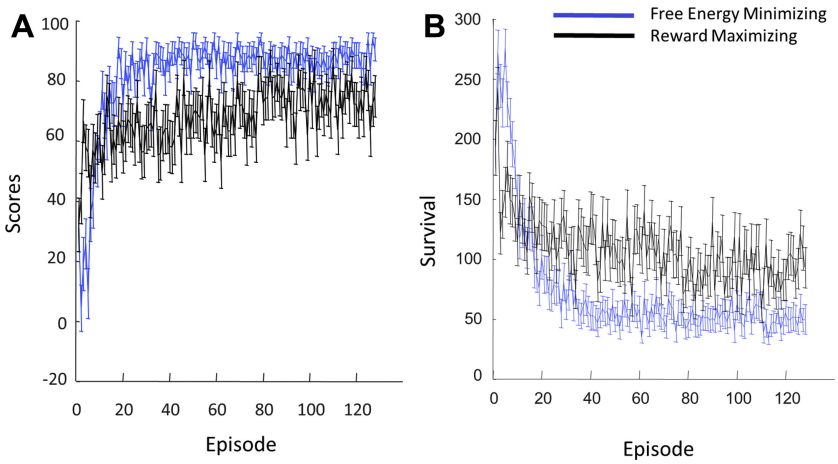

Figure 3: Comparison of free energy and reward-maximizing agents. The plots show mean ± SEM with $p = 0.05$.

uniformity of the transition matrices. In **figure 2C**, it shows reward record from only two simulated instances with different optimization method; now consider both 'reward' and 'survival' (living timesteps of monster) from 50 free energy agents as well as 50 reward-maximizing agents, the results are shown in **figure 3**.

From **figure 3** we could easily find that the average performance of free energy agents is significnatly better than that of the reward-maximizing ones. All the results above demonstrate that active inference outperforms reward-maximizing policies, which is due to the fact of different cost functions. Active inference simultaneously optimize two components, the epistemic value of actions (which reduces uncertainty about state transitions) as well as the extrinsic value of actions. However, only the latter term is optimized under the reward-maximizing policy. In this case, the reward-maximizing agents cannot achieve the desired performance, and for the following subsections, only active inference method is applied.

Active inference simulation vs. humans

Obtaining the best agent mechanism, before further psychiatric analysis, its performance still needs to be assessed with human players. The corresponding comparison of performance is shown in **figure 4**, where alternating trials from the agent's 128 episodes were compared with the human's 64 episodes.

From **figure 4**, the free energy agents are exploring and learning the structure and mechanism of the encironment before exploiting. But the simulated performance mathces human's behavious very fast, only after around 12 actions. As for all the remaining trials, the indistinguishable performance was also retained ($p > 0.05$). In this case, the agents could really achieve similar performance with human in simple tasks.

Complex model (10-state, younger) vs. simple model (6-state, older)

To mimic age-dependent play in the human cohort by agent simulation, firstly, simulate and compare the performances of agents with complex state space (10-state) and simple state space

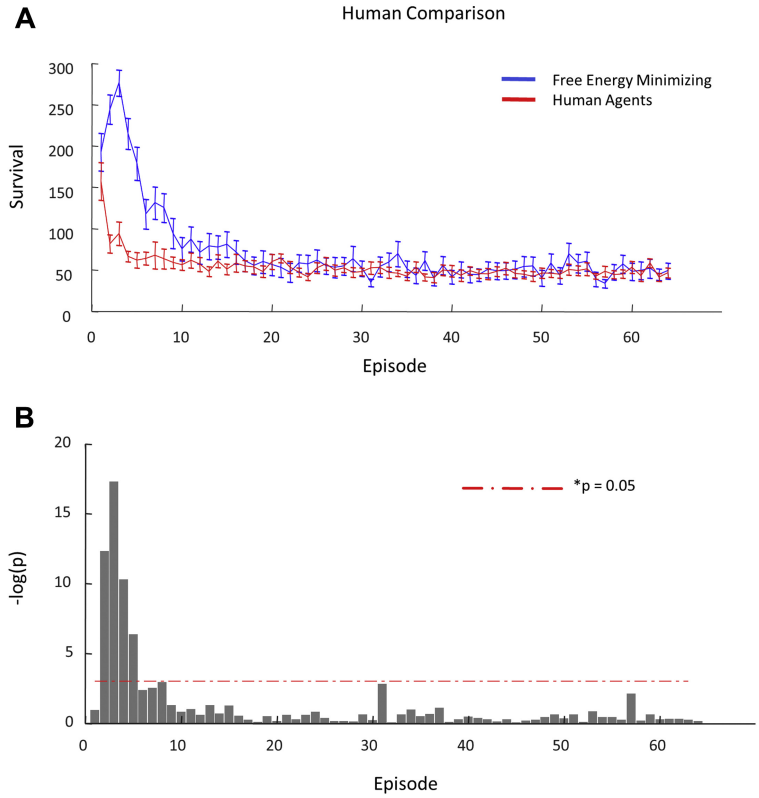

Figure 4: Comparison of free energy agents and human players. **(A)** Average survival scores ± SEM. **(B)** Manhattan plot of statistical difference ($-\log p$-value) for each episode.

(6-state) respectively. Then for human, consider the two groups with different ages separately, too. To quantify the survival metrics recorded from the two groups, the quadratic polynomial curves are used to fit. This is illustrated in **figure 5**.

Over the 64 episodes, the performance of both 6-state and 10-state space models and the older and younger participants share similar features, as well as the second-order quadratic curvatures. In this case, it indeed reflects some similarities of learning efficiencies of the younger with the older and the complex space with the simple one.

Anhedonic priors vs. motivated priors

To simulate the depression feature, a new agent is constructed with a rather flat prior, shown in **figure 6A**. For neuroimaging predictions, the putative neural correlates of activity are simulated in the 'prefrontal cortex' of simulated anhedonic and motivated agents. the amplitude of LFPs proves particular temporal excursion in the motivated compared with anhedonic agents, shown in **figure 6B**. Overall, the LFPs had similar patterns during trials. However, the anhedonic one shows significant change earlier during episodes, and the difference potentials exhibit excursion around episode 20. The side-by-side comparison is replicated for several times, indicating the consistent alteration in state learning process and a concomitant change in LFPs which could be predicted and validated, such as using a time-frequency analysis of electroencephalo-

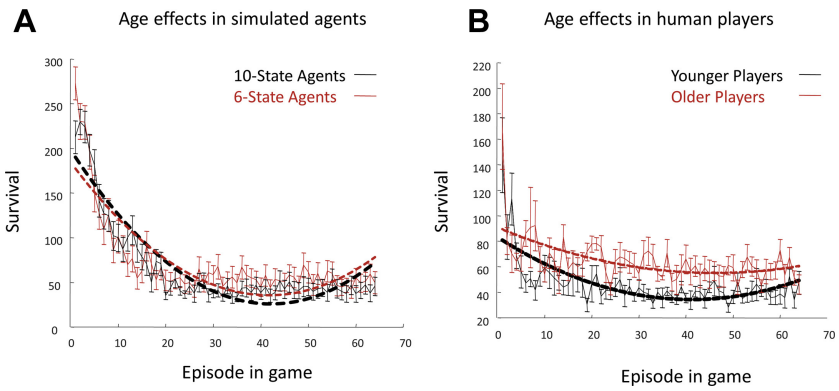

Figure 5: Simulation of aging (mean ± SEM).

gram or magnetoencephalogram. Then these LFPs are adopted to generate BOLD responses within the prefrontal cortex, exhibiting a second small peak at around 22 seconds, which is still consistent with the timing of when the LFP response. Overall, individuals' neural responses could be compared for alternative beliefs, while recapitulating similar forms at the group level.

Discussion

Interests

The paper is interesting and innovative, since it adopts a well-known reinforce learning game platform to analyse psychiatric illness. Also, unlike classic RL optimization target, the active inference with free energy principle shows significant power in simulating human learning and behaviour process, and this makes the simulation and corresponding results more convincing. Meanwhile, the construction of matrix B (state transition probability) is also interesting. In classic RL, the agent usually learn to evaluate the value of state and the value of action at certain state, not trying to know the transition mechanism of the environment itself. In other words, unlike an artificial agent that simply seeks to maximize reward, a free energy-minimizing agent can develop an internal model of environment. Since the active inference comprises both property of normative models and process models, it seems that the active inference converges more closely to the real human brain's learning mechanism. Nevertheless, even in such simple game environment framework, the learning trends that may be reflected in aging could be identified as well as the demonstration of computational phenotypes and neural biomarkers elucidated from game play, with the focus on anhedonic features of depression.

Strengths and weaknesses

Strengths of the paper are considered as follows:

- Unlike simply adopting a classic RL framework to simulate, the paper choose to implement with the active inference and compares the performance with the former one's.
- Adopting the Gym game platform rather than traditional measures or tests. On the game platform, participants would have high engagement as well as interaction, and compliance may be a useful adjunct in early intervention programs.

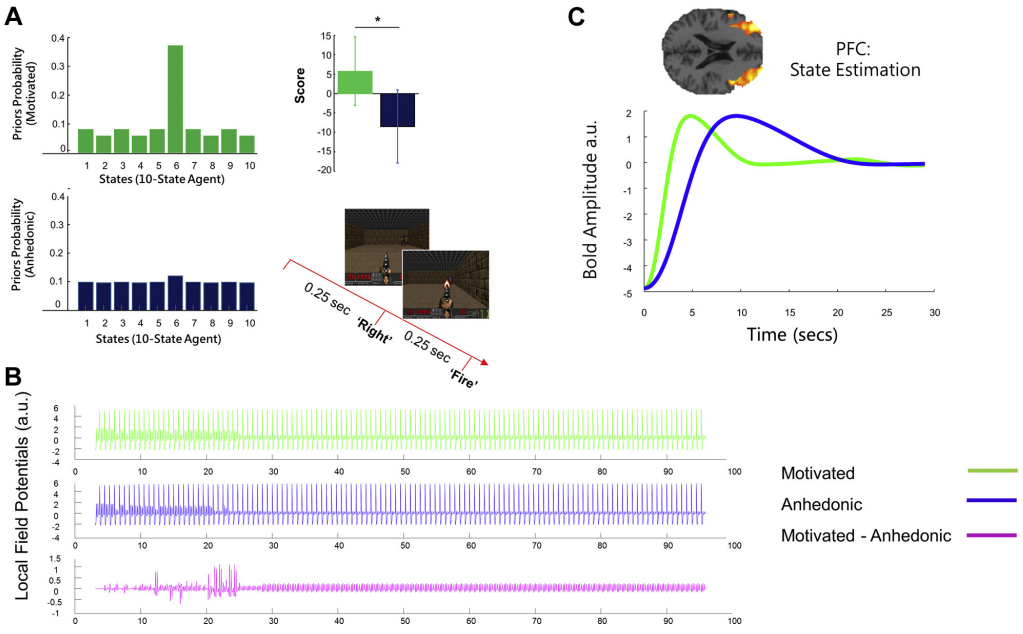

Figure 6: Simulations of anhedonia.

- The agent simulation structure is designed clearly and powerful.

Weaknesses of the paper are considered as follows:

- When simulating the comparison of the different learning mechanism between the younger and the older, 6-state space and 10-state space are adopted respectively. Though the younger has more complex states, the 'middel' state is the same as the older's. In other words, when the agents are simulating the aging, though states classification density is not the same, once it reaches the correct location, the shoot would almost definitely kill the monster. While in reality, the most significant difference during playing such a FPS game between the younger and the older may be the reaction speed, which means that there would be a larger delay of shooting for the older when location is suitable. In this case, I would personally consider to change the time interval of two age groups to represent a slower reaction speed of the older, or change the window width of the 'middle' state, turning it to be a probability of shooting at the monster precisely.

- In the 10-state prior setting, C=$[0.05, 0, 0.05, 0, 0.2, 0.6, 0.05, 0, 0.05, 0]$. Though the prior only slightly affects the final results if convergence is secured (in most cases, it just slows down the learning speed), I would personally consider this as not suitable enough. For instance, C=$[0.04, 0, 0.06, 0, 0.2, 0.6, 0.06, 0, 0.04, 0]$ may be more sensible, indicating different state value between a close relative location and a rather far one.

- In the beginning of human plays, the buttons 'left' 'right' and 'space' are mapped in terms of 'move left', 'move right' or 'shoot' randomly to ensure that human player has no additional prior of the action space, compared with agents (for agents, in the beginning they only know that there are action '0', '1' and '2', not knowing any further function of these actions). However, this only disturbs the human player, since we really know what

the three actions meaning, and only the mapping relation is changed. By this I mean that with additional prior information, people may learn significantly faster and better than the agents. Though this point does not affect the result in this paper, if there is any extension for more complex psychiatric diagnosis, this point should be carefully considered. On the other hand, the buttons are still 'left' 'right' and 'space', so if the direction buttons are changed by the other (e.g., 'left' becomes 'right' and 'right' becomes 'left'), the operation for human then would be much more confusing. I would personally consider to use button '1', '2', '3' to replace the original buttons to avoid extra disturbing.

- In **figure 5**, the paper adopts quadratic polynomial fit, whose function value will decrease and then increase. However, in most common deagnostic plots in machine learning, once the convergence is secured, the curve would decrease and then keep a small fluctuation at certain stable level (just like a white noise with zero mean). In this case, the fitting function form is not suitable in my view. I would personally consider a modified exponential function to fit the plot.

Impact and importance

The paper proposes the use of a principled algorithmic and neurobiological framework for testing hypotheses in psychiatric illness. Though due to the page limitation, it only considered two cases (aging and anhedonia) within one environment (*Doom*), all the results have already revealed the further research potential (active inference on game platform has the potential to be used for hypothesis testing in clinical populations).

Furthermore, since the correlation of simulated agents and human player is demonstrated, in the future, under specific situation, maybe the games could record the operations of the player, as well as some biological signals with special equipment, to capture potential psychiatric illness of the person.

As a dual normative and process theory of the brain, active inference under the free energy principle may be used to reveal structure in behavior and imaging markers in novel experimental settings, allowing clinicians and patients to gain a more comprehensive description, at the algorithmic and mechanistic level, of mental illness.

Potential extensions

- In this paper, the action space and state space are considered to be discrete, due to the simplicity of the game environment itself. However, to mimic the real world condition that human interacts with, more complex game environment should be loaded, and the continuous space could be introduced for futher analysis. In other words, less trivial environments that optimal policy are unknown or difficult to learn could be adopted to compare and examine the performance of human.

- When the paper discusses the rationale of dividing different number of states for matching the groups with different ages, it mentions that, while old people loss adaptive learning ability in brains, younger brains may overlearn unimportant details of the environment's structure. So I'm wondering if it would be interesting to create a 12-state or even 14-state, 16-state of space, to represent the 'overlearn' mechanism. The intuitive idea is that, for an 'underfitting' state space, it only classifying 'left', 'middle', 'right'; for a rather more suitable state space, it consider whether the relative location is 'far left/right' or 'close left/right';

while for an 'overlearn' agent, the 'overfitting' state space may lead to a complex hierarchal classification. This may also have links to certain 'distraction symptoms'.

- In **figure 2C**, around trial 80, there is a decline on performance. The authors explain this point by the failure of Harris corner-detection algorithm. However, since the paper is not focusing on visualization or recognization, the effect of labeling target (the monster in *Doom*) should be removed. In other words, directly representing the locations with the built-in function output by the environment rather than recognizing from image pixels might be a better choice.
- The hypothesis could be tested in certain population of people with anhedonia. For instance, this could be implemented by using estimated prior belief from simulated model to predict optimism scores of individuals.

Acknowledgments and disclosures

This is a Computational Cognitive Neuroscience (CCN, INFR-11036) course essay, which is a report based on published papers in the computational cognitive neuroscience field. Without specific notation, all data, figures and results are directly cited from the original paper, together with corresponding supplement materials.

Reference

Buxton, R. B., Wong, E. C., & Frank, L. R. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. Magnetic resonance in medicine, 39(6), 855-864.

Cullen, M., Davey, B., Friston, K. J., & Moran, R. J. (2018). Active inference in OpenAI Gym: A paradigm for computational investigations into psychiatric illness. Biological psychiatry: cognitive neuroscience and neuroimaging, 3(9), 809-818.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural networks, 15(4-6), 603-616.

Friston, K. (2010). The free-energy principle: a unified brain theory?. Nature reviews neuroscience, 11(2), 127-138.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. Neural computation, 29(1), 1-49.

Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. Computational Psychiatry (Cambridge, Mass.), 1, 2.

Gilbert, J. R., & Moran, R. J. (2016). Inputs to prefrontal cortex support visual recognition in the aging brain. Scientific reports, 6(1), 1-9.

Harris, C., & Stephens, M. (1988, August). A combined corner and edge detector. In Alvey vision conference (Vol. 15, No. 50, pp. 10-5244).

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. Neuron, 80(2), 519-530.

O'doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. Annals of the New York Academy of sciences, 1104(1), 35-53.