ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет информатики, математики и компьютерных наук**

Программа подготовки магистров по направлению

01.04.02 Прикладная математика и информатика

Богатова Маргарита Дмитриевна

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА**

на тему:

**«Исследование построения агентов на основе метода Active Inference»**

**Научный руководитель:**

профессор кафедры прикладной математики и информатики

Крылов Владимир Владимирович

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись

**Рецензент:**

степень

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Подпись

Нижний Новгород  
2022

АННОТАЦИЯ

Определения, обозначения, сокращения

ОГЛАВЛЕНИЕ

[АННОТАЦИЯ 1](#_Toc102039182)

[Определения, обозначения, сокращения 3](#_Toc102039183)

[ОГЛАВЛЕНИЕ 4](#_Toc102039184)

[Введение 5](#_Toc102039185)

[Глава 1. 7](#_Toc102039186)

[Глава 2. 8](#_Toc102039187)

[Глава 3. 9](#_Toc102039188)

[Заключение 10](#_Toc102039189)

[Список ЛИтературы 11](#_Toc102039190)

Введение

Активный вывод — это математическая структура, возникшая в вычислительной нейронауке как теория того, как мозг реализует действие, восприятие и обучение [ссылка на фристона 2016]. В основе активного вывода лежит принцип минимизации свободной энергии, предполагающий в что агенты действуют, противодействуя силам окружающей среды, которые мешают им оставаться в ограниченном наборе желаемых состояний мира, и минимизируют неожиданность сенсорных наблюдений за счет минимизации аппроксимации этой неожиданности, которая называется свободной энергией. Согласно этому предположению, биологические агенты развивают различные навыки, такие как восприятие, действие, планирование и обучение, которые агенты постоянно адаптируют на протяжении всей своей жизни.

Активный вывод и принцип свободной энергии использовались для объяснения и моделирования нескольких сложных процессов в различных дисциплинах. Например, в психологии они использовались для распознавания эмоций [1]. В экономике принцип свободной энергии использовался для формулирования процесса оптимизации агентов с точки зрения их когнитивных искажений[1].

Недавние разработки в области глубокого обучения открыли новые горизонты для изучения и экспериментирования с различными теориями восприятия и поведения. Одним из популярных примеров в этом отношении является обучение с подкреплением (RL) [ 22 ], теория, которая связывает сигналы дофамина в мозге с сигналами вознаграждения, которые можно использовать для подкрепления правильного поведения [ 23 , 24 ], и описывает, как поведение агента может быть оптимизировано через максимизацию вознаграждения [ 25 ]. Сочетание RL с моделями глубокого обучения для оценки функций [ 26 ] позволило достичь успехов в построении искусственных агентов. Такие агенты нашли широкое применение в видеоиграх [26].27 , 28 ], настольных игры [ альфа зиро, мю зиро ] или в управлении роботами [ ddpg, sac ]. Методы глубокого обучения также начинают возникать в контексте активного вывода [variational policy, pmdp, deep ai, control tasks]. Интегрированный с машинным обучением активный вывод называют deep active inference (dAIF).

Задача обучения с подкреплением обычно формулируется как MDP или POMDP, что совпадает с дискретной формулировкой задачи активного вывода. Ввиду этой образующейся связи, активный вывод сейчас часто применятся для решения задач именно в такой формулировке, что позволяет внедрять передовые разработки в области RL в AIF. Как следствие и использовании глубокого обучения, dAIF достигают сравнимых с RL результатов.

Современные задачи управления роботами обычно формулируются как POMDP\MDP и решаются с применением методов обучения с подкреплением. В свою очередь в подавляющем большинстве работ по активному выводу рассматриваются игрушечные дискретные задачи. Вектор исследования в области активного вывода сейчас направлен в сторону расширения области применения на задачи с непрерывным пространством действий и состояний [as reinforsment learning, scaling active inference, contrastive active inference].

На данный момент есть ряд исследований, в которых активный вывод применятся для задач с пространством состояний большой размерности [as pomdp, control, as variation policy], однако исследований для многомерного пространства и состояний, и действий еще очень мало и многие из них не используют все возможности глубокого обучения. Таким образом **задача** адаптации активного вывода для непрерывного пространства действий, решающая **проблему** масштабирования этого метода для комплексных задач таких как управление роботом является **актуальной**.

**Объектом** исследования являются алгоритмы активного вывода, а также методы глубокого обучения для управления роботом. **Предметом** работы является исследование применения теории активного вывода для реальных задач большой размерности. Поэтому, **цель** этой работы состоит в изучении существующих подходов к построению active inference агентов и разработка на их основе алгоритма для решения задачи управления роботом (каким).

Для достижения поставленной цели требуется решить следующие **задачи**:

1.

.

В процессе работы использовались следующие **методы**: анализ литературных источников, описывающих передовые достижения в области … , и обобщение полученной информации; проведение экспериментов по … и сравнение результатов, визуализация и анализ данных результатов проведенных экспериментов.

**Новизна** работы заключается в том, что в ней:

**Практическая значимость**: обученная модель для оценки позы человека может быть использована в других прикладных задачах; разработанный программный модуль может быть использован для разработки приложения, которое упрощает контроль за осанкой человека.

Работа имеет следующую **структуру**:

Глава 1. Обзор литературы и постановка задачи

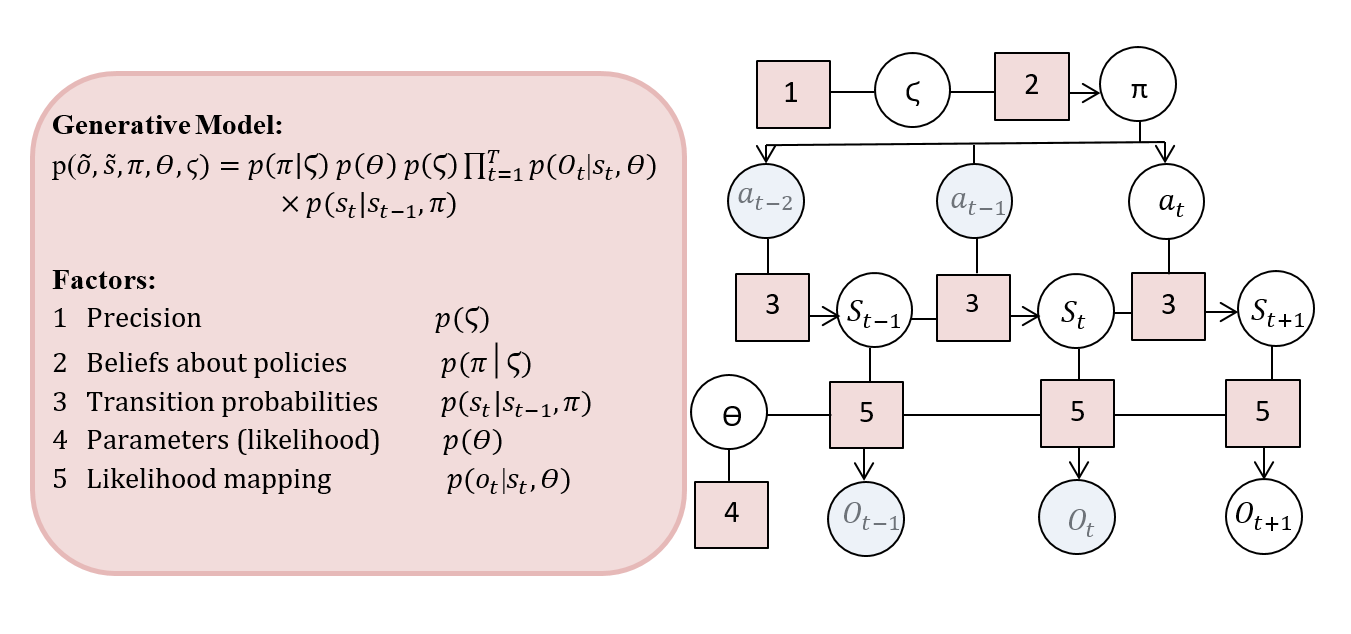
Теория активного обучения была предложена К.Дж. Фристона как единой теории мозга Главная концепция активного обучения состоит в минимизации функционала свободной энергии.

В активном выводе предполагается, что окружение, в котором действует агент описывается как частично наблюдаемый марковский процесс принятия решений(POMDP), which is denoted with the tuple M = {S, A, T, Ω, O, γ}, where S is the set of unobserved states, A is the set of actions, T is the state transition function, also referred to as the dynamics of the environment, Ω is the set observations, O is a set of conditional observation probabilities, and γ is a discount factor (Figure 1). Такая формальность обусловлена тем, что агент не имеет «доступа» к истинному состоянию мира. Используя прошлые наблюдения, агент строит **убеждение** об истинном состоянии.

Агент получает наблюдения из среды, имеющей скрытые состояния, которые агент пытается вывести. Агент также может предпринимать действия, которые изменяют состояние среды и, следовательно, будущие наблюдения.

Поведение агента в среде определятся его генеративной моделью своих наблюдений, состояний и действий, которые можно факторизовать следующим образом:

p(o, s, a, γ)



Путем оптимизации генеративной модели агент улучшает свою способность выяснить, какое действие привело к наблюдению. Каждый раз, когда агент вычисляет ожидаемую свободную энергию (EFE) каждой политики, которая количественно определяет вариационную свободную энергию (VFE) и изменяет среду, чтобы свести к минимуму VFE.

Вариационная свободная энергия

Ключевой смысл принципа свободной энергии состоит в том, что агент стремится найти равновесие между своей внутренней моделью и внешним миром. В терминах Баесовского вывода, агент максимизирует обоснованность (evidence) модели, приближая внутреннее состояние, описываемое плотностью q(s), к апостериорному миру p(s|o). Помимо интуиции, введение вариационного распределения необходимо по причине того, что p(s|o) в большинстве реальных случаев нельзя посчитать точно.

Для того, чтобы подобрать лучшую аппроксимацию q(s) минимизируют дивергенцию Кульпака-Лейблера (KL) между распределениями q(s) и p(s|o). Для простоты записи, рассмотрим упрощенную генеративную модель .

Из этого определения вытекает формула вариационной свободной энергии (VFE):

Эта формула представляет собой функцию внутренних убеждений агента об окружающей среде, и означает верхнюю границу неожиданности наблюдений. VFE измеряет меру расхождения между вариационной плотностью q(s) и генеративной моделью p(s,o).

Далее, формула свободной энергии может быть переписана следующим образом:

С одной стороны, минимизация свободной энергии подразумевает максимизацию точности (второе уравнение, второе слагаемое) при одновременном снижении сложности апостериорного распределения (второе уравнение, первое слагаемое). С другой стороны, это подразумевает оптимизацию обоснованности модели (первое уравнение).

Очередной раз преобразуем формулу: используем условие марковского процесса о том, что распределение вероятностей будущих состояний процесса зависит только от нынешнего состояния и заменим на .

Это наиболее часто используемый в литературе вид формулы свободной энергии для урезанной генеративной модели. В статье [var policy] была выведена формула для генеративной модели изображенной на схеме[] которая берет во внимание распределения действий агента. Выглядит она следующим образом:

Ожидаемая свободная энергия

«Активной» составляющей агента активного вывода является взаимодействие с окружением. Теперь необходимо взять во внимание то, что агент совершает действия. Для того, чтобы действовать в среде наиболее оптимально, в подходе предлагается помимо минимизации вариационной свободной энергии, также минимизировать ожидаемую свободную энергию (EFE), которая представляет собой комбинацию свободных энергий всех ожидаемых траекторий в будущем с учетом текущего состояния и действия.

Для вывода формулы рассмотрим генеративную модель . Наиболее критичным для определения распределением является которое означает предпочтения агента, или другими словами наблюдения, которые агент ожидает получить.

*=*

Интуитивно, выведенная формула означает, что агент будет выбирать такие действия, для которых наблюдения содержат больше информации о состояниях, которые будут вести к получению предпочтительных результатов и в то же время сохраняя действия настолько вариативными, насколько это возможно.

Таким образом, мы определили приближенные формулы, которые обычно являются основой для различных подходов. Так, когда говорят об активном выводе, обычно оперируют следующими понятиями:

* – модель переходов (*transition model*).
* – предпочтения агента *(agent preferences)*.
* - условное распределение действий, обычно в литературе называет моделью действий.
* – модель наблюдений *(observation model)*. Также в литературе называется репрезентацией.

Связь с обучением c подкреплением

Активный вывод был предложен как альтернатива обучению с подкреплением. Эти подходы призваны решать одни и те же задачи, и за частую оперируют одними и теми же формулами, однако идеи, лежащие в основе, у них разные. Например, концепция «вознаграждения» в активном выводе сильно отличается от вознаграждения в RL, поскольку вознаграждение — это не сигналы, используемые для привлечения агента, а скорее сенсорные состояния, которые агенты стремятся часто посещать, чтобы минимизировать свою свободную энергию. Авторы активного обучения утверждают, что агенты здесь функция награждения совсем не обязательна и агенты смогут изучать свою собственную.

Активный вывод в задачах непрерывного управления

Фокус этой работы направлен на применение активного вывода для решения задачи управления. Особенности этой задачи состоит в том, что при ее рассмотрении как POMDP, внутреннее состояние системы и действия задаются следующим образом: и .

Реализация агента активного вывода, который бы действовал в окружении с непрерывным пространством состояний и действий на текущий момент открытая проблема. Потому что формулы для расчета и обучения aif агента применимы только для дискретных случаев. Так, в [devistified and compared] плотности в уравнении свободной энергии задаются матрицами. Например, кодируется набором матриц для каждого возможного действия, строки и столбцы которых задают плотность для каждой возможной пары состояний. Такой подход называют **табличным активным выводом**. Очевидно, в непрерывной среде такие вычисления будет проводить очень проблематично.

Еще более сложная ситуация с непрерывным пространством действий. Чтобы сделать вывод о своей политике, агент должен вычислить ожидаемую свободную энергию (EFE) для каждой политики. На практике это означает, что для *каждой* *возможной* *политики* агенту необходимо на определенный горизонт времени вперед сгенерировать состояния и наблюдения, в которых он будет, если будет следовать этой политике. И далее вычисляется свободная энергия этих состояний и оценивается ожидаемая свободная энергия для каждой политики. Когда доступных политик и действий неопределенно много, такой способ подсчета ожидаемой свободной энергии также становится неприменим.

Из-за этих проблем с масштабированием табличный активный вывод применялся в основном к игровым задачам, с небольшими пространствами состояний или действий.

Глубокий активный вывод (dAIF)

Во многих недавних научных работах были предприняты попытки масштабирования активного вывода для более комплексных и реальных задач. В последние годы для построения mdp/pomdp агентов большую популярность имеют методы глубокого обучения с подкреплением. Именно применение нейросетей позволило достигнуть значительных успехов в этой области. Эта тенденция не обошла стороной и теорию активного вывода - использования нейронных сетей для аппроксимации плотностей положило начало развитию активного обучения в этом направлении.

Насколько это известно, [4] является первой работой, в которой предлагается аппроксимировать модели наблюдений и переходов с помощью глубоких нейронных сетей. Авторы представляют все плотности в виде диагонального гаусовского распределения, а для аппроксимации параметров этих распределений используют однослойные сети с шестнадцатью нейронами. Также именно они ввели понятие *deep active inference* или *dAIF.*

Большое влияние на область dAIF также оказала работа [5]. Авторы использовали последние на то время разработки в области обучения с подкреплением и создали некоторый смешанный алгоритм. Основываясь на этом подходе были разработаны такие алгоритмы как [6] и [7]. Вкратце, авторы аппроксимировали с помощью нейронных сетей модели наблюдений и переходов, но что самое интересное, позволили агенту самому предсказывать значение ожидаемой свободной энергии для сразу всех доступных действий, и сделано это тем же способом, как это сделано в обучении с подкреплением при аппроксимации Q-функции.

Возвращаясь к проблеме масштабирования активного вывода, далее статьи будут рассматриваться с точки зрения того, какие именно методы предлагаются и предлагаются ли вообще для работы с непрерывными пространствами состояний и действий.

Наиболее интересной в этом плане является работа [8], в которой авторы преследуют такие же цели что и в это исследование. Все плотности они рассматривают как нормальные распределения, а их параметры аппроксимируют с помощью нейронных сетей. Модель наблюдений строится на основе вариационного автоэнкодера, что позволяет применять метод в случае высокоразмерных наблюдений (например, изображений). Ключевой особенностью здесь является то, как реализован выбор политики и планирование. В каждый момент времени они выбирают такое действие, которое будет математическим ожиданием модели действий, которая также представлена как нормальное распределение. Для это из нее семплируются конечная последовательнось действий, для каждой такой последовательности производится оценка свободной энергии и далее модель действий оптимизируется с помощью метода кросс-энтропии (CEM) [9]. Авторы показывают, что такой вид dAIF является конкурентным с передовыми методами. Некоторым продолжением и в своем роде упрощением предыдущего подхода стала работа [10], в которой авторы попробовали усилить алгоритмы обучения с подкреплением с помощью идей активного обучения. На основе принципа ожидаемой свободной энергии была разработана целевая функция, минимизируя которую мы могли бы получить оптимальную модель действий. Благодаря такому упрощению, этот подход можно использовать и для задач с большим входом и непрерывными действиями.

В исследовании [11] рассматривается отличный от стандартного глубокого обучения подход к выбору оптимального действия (то есть минимизации ожидаемой свободной энергии в пределах некоторого горизонта времени). Авторы предлагают использовать MCTS для выбора оптимальной траектории в дереве решений, которое образуется, если рассматривать задачу минимизации ожидаемой свободной энергии для конечного числа возможных действий.

Наконец, из наиболее релевантных последних исследований выделяется [12], авторы которого адаптировали стандартный алгоритм обучения с подкреплением - актер-критик - для предсказания действий, и тем самым создали модель действий, которая применима для непрерывного пространства. Помимо этого, было предложено использовать так называемое контрастное обучение. Как результат, алгоритм, который работает с непрерывным пространством состояний и действий, превосходящий другие dAIF методы на дискретных задачах.

Выводы

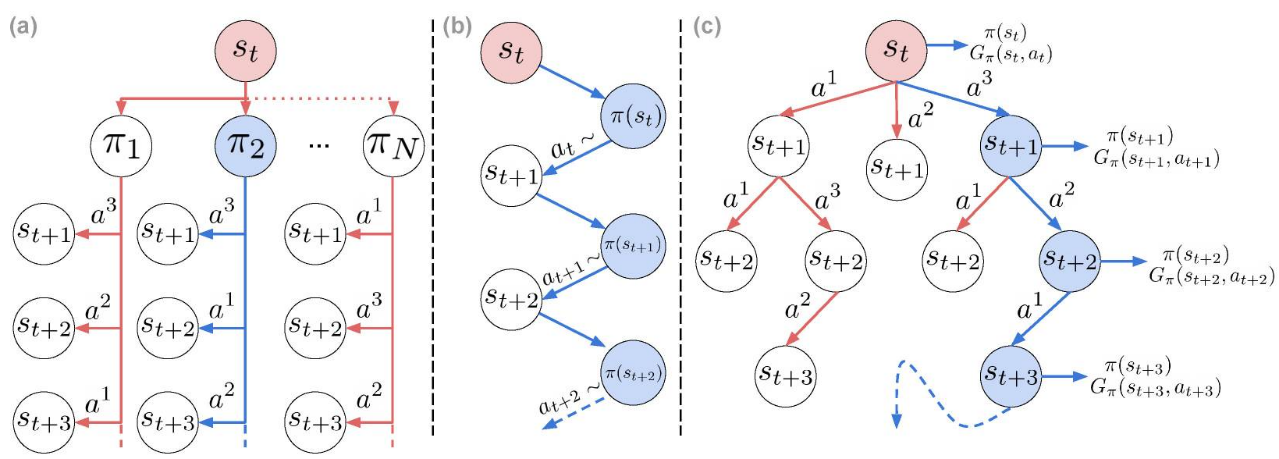


Рисунок 2. Различные подходы для выбора оптимального действия. Источник [1].

Как скалируют наблюдения

Как скалирует дейсвтия и производят планирование

Какие проблемы у текущих подходов - лишь единичные статьи рассматривают интеграцию нейронных сетей и роботов, так как в последних пространство действий непрерывное.

На данный момент существует лишь несколько единиц работ, но они поялвля.тся каждый день.

Одна работа существует но отходит от концепуии активного вывода, а другая появилась параллельно с написанием текущей.

Однако задачи обучения с подкреплением на практике часто являются имеют многомерные пространства состояний и/или непрерывные пространства действий. На данный момент, все предложенные daif подходы применяются только для задач с ограниченным числом действий[],

Примеры существующих решений этой проблемы в аи, ее результаты и недостатки.

Существующие решения в целом в реинфорсмент лернинг

Одно из отличий pomdp и scaling active inference – это то что они исключают часть связанную с q(a|s) из уравнения свободной энергии. И его оптимизируют уже через метод кросс энтропии или msts

~~AIF возникла в конце 2000-х годов как единая теория функций мозга [3], и с тех пор исследователи пытаются внедрить эту теорию в различные подходы для решения задачи обучения с подкреплением. В частности, особым успехом пользуются подходы, которые интегрируют нейронные сети и активный вывод [ вот они слева направо… ]. Такие подходы классифицируют как deep active inference (dAIF).~~

Структура активного вывода и, в частности, ее недавняя формулировка как частично наблюдаемый марковский процесс принятия решений (POMDP) ​​в последние годы приобрели все большую популярность как полезный подход к моделированию нейрокогнитивных процессов. Эта структура является очень общей и гибкой, поскольку ее можно настроить для моделирования любого когнитивного процесса, а также для имитации прогнозируемых нейронных ответов на основе сопутствующей теории нейронных процессов

AIF является интересной структурой, поскольку она объединяет оценку состояния, управление и изучение модели мира как процессы вывода, которые решаются путем оптимизации единственного целевого функционала - свободной энергии - что очень схоже с вариационным байесовским выводом. (актуальность?). Исследования показывают, что активный вывод работает, по крайней мере, так же хорошо в простых средах [variational policy, Активный вывод OpenAI ] и лучше в средах, характеризующихся изменчивостью, неоднозначностью и чувствительностью к контексту [as pomdp ].

Глава 2. теоретическое Описание подхода

In the current paper, we take an alternative approach and employ amortized inference [24], which utilizes function approximators (i.e., neural networks) to parameterize distributions. Free energy is then minimized with respect to the parameters of the function approximators, and not the variational parameters themselves

In our model we do not use this approach, but instead use a learned amortized inference distribution Q(a|s) and minimize this using a variational approach on the divergence with the approximated true posterior of the value function p(a|s),which is learned through a bootstrapping estimation procedure. Due to this our method is end-to-end differentiable and all networks can be trained through gradient descent on the variational free-energy.

Глава 3. реализация

In Reinforcement Learning, you really shouldn't typically be paying attention to the precise values of your loss values. They are not informative in the same sense that they would be in, for example, supervised learning. The loss values should only be used to compute the correct updates for your RL approach, but they do not actually give you any real indication of how well or poorly you are doing.

**This is because in RL, your learning targets are often non-stationary; they are often a function of the policy that you are modifying (hopefully improving!)**. It's very well possible that, as the performance of your RL agent improves, your loss actually increases. Due to its improvement, it may discover new parts of its search space which lead to new target values that your agent was previously completely oblivious to.

Your only really reliable metric for how well your agent is doing is the returns it collects in evaluation runs.

Заключение

Список ЛИтературы

1. Pietro Mazzaglia, Tim Verbelen , Ozan Çatal and Bart Dhoedt, The Free Energy Principle for Perception and Action: A Deep Learning Perspective. 2022.
2. Pablo Lanillos, Cristian Meo, Corrado Pezzato, Ajith Anil Meera, Mohamed Baioumy, Wataru Ohata, Alexander Tschantz, Beren Millidge, Martijn Wisse, Christopher L. Buckley, and Jun Tani. Active Inference in Robotics and Artificial Agents: Survey and Challenges. 2021.
3. How Active Inference Could Help Revolutionise Robotics 2022
4. Ueltzhöffer K. Deep active inference. Biol. Cybern. 2018;112:547–573. doi: 10.1007/s00422-018-0785-7
5. Millidge B. Deep Active Inference as Variational Policy Gradients. arXiv. 2019 doi: 10.1016
6. Aleksey Zelenov, Vladimir Krylov, Deep active inference in control tasks, 2021.
7. Otto van der Himst, Pablo Lanillos, Deep Active Inference for Partially Observable MDP
8. Alexander Tschantz, Manuel Baltieri, Anil. K. Seth, Christopher L. Buckley. Scaling active inference. 2019
9. Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. arXiv:1811.04551 [cs, stat], 2018
10. Alexander Tschantz, Beren Millidge, Anil K. Seth, Christopher L. Buckley. Reinforcement Learning through Active Inference 2020
11. Zafeirios Fountas, Noor Sajid, Pedro A.M. Mediano, Karl Friston. Deep active inference agents using Monte-Carlo methods, 2020
12. Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Contrastive Active Inference. 2022