Нейробиологическая теория Карла Фристона: критический обзор

А. В. Гришаев[1] В. Ф. Сазонов[2]

В данной работе впервые в русскоязычной литературе излагается нейробиологическая версия теории Карла Фристона. Впервые эта теория излагается целостно, подробно и логически выстроено в рамках одной статьи и, насколько это возможно, в адаптированном для понимания нейробиологов и нейрофизиологов виде. Обсуждается то, каким образом Фристон применяет эту теорию на практике. Проводится критический анализ внутренних проблем и противоречий в теории Фристона.

Ключевые слова: принцип свободной энергии, гипотеза байесовского мозга, предиктивное кодирование, ошибка предсказания, наблюдаемые состояния, ненаблюдаемые состояния.

1. Введение

Одной из важных тенденций в современной нейробиологии и нейрофизиологии является развитие представлений о прогностической (предиктивной) деятельности нервной системы. Возможно, что именно «предсказательная» деятельность как раз и является основной задачей нервной системы, обеспечивающей её выживание как системы, а заодно и выживание всего организма. Одна из интереснейших, но небесспорных, разработок в этом направлении принадлежит широко известному в качестве одного из изобретателей воксельной морфометрии (Voxel-based morphometry) британскому нейробиологу Карлу Фристону (K.J. Friston).

Примерно полтора десятилетия назад К. Фристон предложил амбициозную теорию (для краткости мы далее будем говорить просто «*теория Фристона*»), которая, согласно заявлениям автора [1]-[12], объединяя в рамках единого подхода экспериментальные данные о восприятии,

[1] *Гришаев Александр Викторович* — выпускник каф. биологии и методики её преподавания Института естественных наук РГУ им. С. А. Есенина, e-mail: ecology.ag@yandex.ru.

Grishaev Alexander Victorovich — post-graduate student, Ryazan State University named after S.A. Esenin, Institute of Natural Sciences, Department of Biology and Its Teaching Methods.

[2] *Сазонов Вячеслав Федорович* — к.б.н., доцент каф. биологии и методики её преподавания Института естественных наук РГУ им. С. А. Есенина, e-mail: kineziolog@mail.ru.

Sazonov Vyacheslav Fedorovich — Ph. D., Associate Professor, Ryazan State University named after S.A. Esenin, Institute of Natural Sciences, Department of Biology and Its Teaching Methods.

обучении и двигательной активности, предлагает целостную теоретико-информационную и математическую конструкцию для объяснения самых разнообразных аспектов работы мозга.

Истоки этой теории восходят к представлениям ученого-универсала Г. фон Гельмгольца [13]. Гельмгольц в работе «Von den wahr-nehmungen im Allgemeinen» из третьего тома его «*Handbuch der physiologi-schen Optik*» [13] предложил идею о восприятии как о «бессознательном выводе» или «бессознательном заключении» (на немецком: unbewusster Schluss, или на английском: unconscious inference [14], unconscious conclusi-on [15]). Сущность этой идеи состоит в том, что восприятие представляет собой процесс вывода, аналогичный логическому выводу (то есть выводу в логике), но только носящий бессознательный характер, в то время как логический вывод обычно рассматривают как сознательный процесс. Проще говоря, на основе исходных бессознательных суждений о мире (в данном случае это сенсорная информация, а также прошлый опыт субъекта, его знания об окружающем мире и т. д.) через бессознательные умозаключения осуществляется переход к новым суждениям — бессознательным заключениям о мире— ощущениям. Однако и у идей Гельмгольца, и, соответственно, теории Фристона, а также у других сходных теорий [16], [17] с общим названием «Predictive processing», как недавно выяснилось благодаря работе Л. Свонсона [18], есть влиятельнейший предтеча — И. Кант с его трансцендентальной философией, где центральное в этой философии понятие — «вещь в себе» (на немецком: Ding an sich). Кроме того, идеи, аналогичные по смыслу идее об «бессознательном выводе», выдвигались Ф. Бэконом [19], Т. Гоббсом и Ф. Нортом [20].

Сама идея о восприятии как о бессознательном выводе имела и имеет как сторонников [21]-[26], так и противников [14], [27], [28]. В данной статье мы не будем обсуждать аргументы в поддержку этой идеи или её критику, так как это является темой для отдельного большого обсуждения. Заинтересованного читателя отсылаем к работе Г. Хатфилда [27], где как раз и содержится подробное обсуждение этой идеи, а также аргументы «за» и «против» неё.

Итак, идеи Гельмгольца о восприятии послужили отправной точкой для последовавших затем интерпретаций и наполнения этих представлений новыми смыслами.

В 1980 году Р. Грегори (известен среди отечественных ученых в первую очередь своей книгой «*Разумный глаз*») [21], продолжая идею «бессознательного вывода», сравнил перцепции с гипотезами в науке. А сам процесс восприятия он сопоставил с деятельностью ученого-экспериментатора, выдвигающего гипотезы, а затем проверяющего их в эксперименте.

В 1990-е годы идеи Гельмгольца нашли свое продолжение в работе Дж. Хинтона с соавторами [29], выдающегося специалиста в области искусственного интеллекта и машинного обучения. Авторы предложили считать, что «перцептивная система представляет собой устройство для статистического вывода, чья функция заключается в выводе вероятных причин сенсорного входа». Кроме того, они использовали в этой работе такие понятия (похожие по содержанию на понятия, используемые Фристоном), как «модель распознавания» и «порождающая модель». Модель распознавания, согласно Хинтону и соавторам, используется для «вывода вероятностного распределения основных причин из сенсорного входа», а порождающая модель используется для «обучения модели распознавания». Порождающая модель также обучается. Иначе говоря, *Хинтон и соавторы наполняют* исходное довольно расплывчатое понятие Г. фон Гельмгольца *«бессознательный вывод» статистическим смыслом*, значительно расширяя и уточняя это понятие.

Другой важной работой, которую можно считать предтечей теории Фристона, является работа Рао и Балларда 1999 года о предиктивном кодировании (predictive coding), описывающем процесс обработки зрительной информации [30]. Заметим, что в отечественной технической литературе и словарях термин predictive coding присутствует уже достаточно давно и переводится либо как «кодирование с предсказанием» [31], [32], либо как «предиктивное кодирование» [32]. Мы будем использовать в дальнейшем именно термин «предиктивное кодирование».

Идея *предиктивного кодирования* (по отношению к зрительной системе) состоит в том, что обратные связи от вышестоящих в иерархии областей зрительной коры несут прогнозы (предсказания) нейронной активности для нижестоящих областей зрительной коры, а прямые связи от нижестоящих областей зрительной коры возвращают вышестоящим областям ошибки между прогнозами (предсказаниями) и фактической активностью нижестоящих областей коры [30]. В дальнейшем, в работах других авторов [33]-[35], и самого Фристона [1], [5], идея предиктивного кодирования стала распространяться не только на обработку зрительной информации, но также и слуховой, и, в целом, на работу всей коры.

Стоит заметить, что несмотря на сходство этой идеи с нервной моделью стимула Е. Н. Соколова [36], состоящее в том, что в обоих случаях сравнивается степень сходства двух потоков информации и выдается «сигнал рассогласования» при несовпадении этих потоков, эта идея имеет существенные отличия. Предиктивное кодирование в формулировке Рао и Балларда [30] совершенно не связано с ориентировочной реакцией, в отличие от нервной модели стимула. Оно связано с потоком информации по иерархии зрительной системы. Нервная модель стимула ничего подобного не содержит. Нервная модель стимула формируется, согласно

Соколову [36], на нейронах гиппокампа, в то время как исходная формулировка предиктивного кодирования вообще не связана с работой гиппокампа. В версиях предиктивного кодирования, как и теория Фристона [37],выходящих за рамки зрительной системы, и включающих себя самые разнообразные области мозга, гиппокамп может являться лишь одним из многих уровней иерархии, в котором сравниваются потоки информации. Существенно различается и то, что сравнивается. Согласно предиктивному кодированию на самом нижнем уровне иерархии сравнивается предсказание, приходящее с вышележащего уровня иерархии, с сенсорной информацией; на всех же остальных уровнях иерархии уже сравниваются предсказания, приходящие с вышележащих уровней иерархии, с ошибками предсказания, приходящими из нижележащих уровней иерархии. Нервная модель стимула не содержит таких понятий как «предсказание» и «ошибка предсказания», и сравнивается в данном случае не предсказание с его ошибкой, а сенсорная информация с её моделью.

Из самой идеи предиктивного кодирования, по нашему мнению, прямо не следует наличие в ней какой-то вероятностной составляющей, однако эта гипотеза может быть трактована в вероятностном смысле, где прогнозы (предсказания) становятся *вероятностными* прогнозами. Помимо этого, тот математический аппарат, который используют Рао и Баллард, содержит вероятностную составляющую [30], что дополнительно указывает на возможность вероятностной трактовки этой гипотезы.

Ещё одной важной для становления теории Фристона идеей является гипотеза байесовского мозга или гипотеза байесовского кодирования [38].

Согласно *гипотезе байесовского кодирования* мозг представляет сенсорную информацию в форме вероятностных распределений, кодируя и вычисляя оптимальным, по Байесу, образом функции плотности вероятности или приближения к функциям плотности вероятности [38]. Стоит заметить, что представления Хинтона [29] и гипотеза байесовского кодирования близки по смыслу и взаимно дополняют друг друга. Подробнее о гипотезе байесовского кодирования можно узнать в обзорной статье D.C. Knill и A. Pouget [38] или в списке литературы этой статьи.

Таким образом, творчески осмысливая идеи Гельмгольца, Хинтон и другие исследователи, начавшие применение байесовских методов для объяснения тех или иных аспектов восприятия, заложили основу для создания Фристоном его теории.

Мы, в свою очередь, считаем важным моментом акцентировать внимание читателя на том, что в теории Фристона используется именно *вероятностный (байесовский) подход* к объяснению и моделированию нервных процессов.

Теория Фристона была создана не единомоментно, общее здание теории создавалось и описывалось её автором во множестве статей. Тем не

менее, развитие теории не закончилось и продолжается в течение уже более полутора десятилетий, начавшись от формулирования основных принципов в первых статьях [5], [11], затем обогатившись основами математического аппарата в формулировке с непрерывными величинами [2], [3], [7], [8], в дальнейшем развившись в концепцию активного вывода [4], [6], и в последующих статьях развиваясь, главным образом, в формулировке с дискретными величинами [12], [39].

Математический аппарат теории Фристона имеет лишь небольшие отличия от математического аппарата байесовских методов глубокого обучения (deep learning). Любопытно, что математический аппарат, изначально разрабатываемый и применяемый в искусственном интеллекте (ИИ), применяется Фристоном, в несколько видоизмененном виде, к «интеллекту естественному». Не менее интересно и то, что совсем недавно случился и обратный поворот, когда в нескольких работах [40], [41] математический аппарат теории Фристона применили уже для задач ИИ.

Научную значимость теории Фристона подчеркивает огромная цитируемость его работ, посвященных этой теории. Так, согласно информационно-поисковой системе Google Scholar, 5 самых цитируемых статей с изложением теории Фристона имеют цитируемость $\sim$ 1000-5000 цитирований [42].

Однако изучение и понимание теории Фристона, на наш взгляд, существенно осложняется тем, что разные аспекты этой теории разбросаны в десятках статей автора, и его теория на данной момент не представлена в каком-то едином обобщающем научном труде в целостном виде во всех своих аспектах и с подробным и последовательным изложением. К тому же в разных статьях Фристона основные математические тождества этой теории даются с разными обозначениями, и поэтому не всегда ясно то, откуда взялось какое-то математическое тождество, и как математические тождества из разных статей соотносятся друг с другом, что затрудняет понимание теории Фристона. Например, разными символами (буквами) в разных работах обозначены: сенсорные данные, состояния среды и состояния мозга [5], [6], [43], разные виды информационных энтропий и информационных энергий [5], [6], [11], [43], [44].

Кроме того во многих работах [45]-[48], которые обращены к теории Фристона, но в которых Фристон не является автором или членом авторского коллектива, рассматриваются только отдельные аспекты этой теории, и, как правило, наиболее простые для понимания идеи, например, связанные с предиктивным кодированием. Однако более сложные для понимания вещи, такие как, например, скрытые состояния, разные виды информационных энтропий и энергий и т.д., а также целостный взгляд на всю теорию – как правило, остаются вне рассмотрения.

Помимо этого мы считаем, что существует ещё одна проблема с пониманием теории Фристона, которая заключается в том, что она изложена на языке байесовской теории вероятностей, а для большинства нейробиологов (то есть для тех людей, кому она в первую очередь интересна и полезна) этот язык является малознакомым. Полноценное понимание теории Фристона требует от читателя математической подготовки. Так, от читателя требуются *как минимум* элементарные знания в байесовской теории вероятностей, байесовских методах машинного обучения, теории информации, математическом анализе, численных методах решения дифференциальных уравнений и т.д. Поэтому глубокое и всестороннее понимание этой теории вызывает вполне понятные затруднения у нейробиологов. Стоит также заметить, что подходящих книг с систематическим и подробным изложением байесовской теории вероятностей на русском языке нами обнаружено не было, что ещё более затрудняет понимание теории Фристона отечественными нейробиологами. Кроме того, байесовская теория вероятностей имеет свою специфическую терминологию, и многие термины из этой теории могут сбивать с толку неспециалистов, вызывая у них ложные ассоциации и интерпретации.

В связи с вышеизложенными обстоятельствами ***актуальным*** является критическое изложение теории Фристона, насколько это возможно, в максимально понятной (в том числе для российских нейробиологов), целостной и последовательной форме в одной статье.

Новизна данной статьи. Насколько нам известно, данная работа является первой публикацией на русском языке, содержащей подробный обзор и критический анализ нейробиологической теории К. Фристона.

Целями данной работы являются:

- целостное и последовательное описание нейробиологической теории К. Фристона на основе обзора его работ, адаптированное для понимания нейробиологов и нейрофизиологов;
- критический анализ внутренних проблем и противоречий в теории Фристона, в том числе и математических;
- обсуждение возможностей и проблем применения этой теории для моделирования более приближенных к реальности нейрональных процессов по сравнению с теми, которые моделируются в работах Фристона на данный момент.

Несмотря на то, что согласно Фристону, его теория имеет общебиологический характер [43], то есть обладает способностью описывать не только работу мозга, но и работу других биологических систем, мы в нашей работе ограничимся *нейробиологическим направлением* его теории,

на основании того, что последнее наиболее полно разработано и наиболее детально описано в работах Фристона.

Мы также обращаем внимание на то, что теория Фристона представлена в его работах в двух вариантах: формулировка с использованием дифференциальных уравнений, непрерывных значений времени и случайных величин (в дальнейшем «непрерывная формулировка») и формулировка, основанная на теории марковских процессов принятия решений и дискретных значениях случайных величин. Обе формулировки математически примерно эквиваленты, однако непрерывная формулировка теории описана в работах Фристона наиболее подробно, к тому же, и об этом пишет сам Фристон [12], теория в непрерывной формулировке более биологически реалистична, поэтому в нашей работе мы будем описывать теорию Фристона в *непрерывной формулировке*.

Итак, мы выделим и последовательно рассмотрим следующие основные положения, на которых базируется нейробиологическая теория К. Фристона:

- Принцип свободной энергии.
- Фристоновская формулировка гипотезы байесовского мозга.
- Фристоновская формулировка предиктивного кодирования.

Следует заметить, что словосочетание «фристоновская формулировка» означает, что гипотезы байесовского мозга и предиктивного кодирования у Фристона имеют смысловое наполнение, несколько отличное о того, с каким они были изначально предложены другими авторами [30], [38]. Поэтому читателям желающим более подробно познакомится с этими гипотезами, мы рекомендуем начать с исходных статей по этим гипотезам [30], [38].

Согласно собственным представлениям Фристона [1], наиболее общей идеей из вышеперечисленных является *принцип свободной энергии*. Поэтому мы начнем изложение теории Фристона именно с него.

Ещё раз отметим, что поскольку в работах Фристона не соблюдается изложение его теории с логически выстроенным и последовательным формулированием основных понятий и выведением основных положений, то мы в нашей работе предпримем попытку самостоятельно произвести такое логически выстроенное и последовательное изложение. По ходу такого изложения мы будем вводить дополнительные важные предположения, фигурирующие в теории Фристона, но прямо не относящиеся к этим трём положениям. На этом будут сделаны отдельные акценты для каждого вновь вводимого предположения.

2. Принцип свободной энергии

Сначала мы приведем аргументацию в поддержку принципа свободной энергии, которую приводит Фристон, и наши замечания по этому поводу. А затем постепенно мы будем разъяснять содержание этого принципа и смежные вопросы.

В мотивировке принципа свободной энергии Фристон пишет о том, что биологические системы функционируют в условиях постоянно меняющейся среды, испытывают её воздействие на себя и с помощью гомеостаза поддерживают своё внутреннее состояние (или состояния) в пределах определённых границ [1], [5], [11], [49]. И эти тезисы, на первый взгляд, не вызывают вопросов.

Однако заметим, что уже в самом начале мотивировки своего принципа свободной энергии Фристон не указывает, что же имеется ввиду под словом «состояния». Не находится определения этого термина и в процессе анализа всех рассмотренных нами работ Фристона. Поэтому нам приходится излагать и рассматривать теорию Фристона в том виде, в каком она была сформулирована самим автором — без чёткого указания, что же подразумевается под термином «состояние». Хотя этот термин, на наш взгляд, требует подробного разъяснения.

Затем Фристон, например, в работе [1] продолжает рассуждения следующим образом: биологические системы «*противостоят естественной тенденции к беспорядку*», и далее через несколько предложений: «*... это поддержание порядка... отличает биологические системы от самоорганизующихся систем; действительно физиология биологических систем может быть почти полностью сведена к их гомеостазу*». Логично предположить, что здесь Фристон ведёт речь о термодинамической энтропии. И биологические системы противостоят естественной тенденции к возрастанию термодинамической энтропии и являются, таким образом, термодинамически негэнтропийными системами. Такое представление о биологических системах, как о термодинамически негэнтропийных, не является новым и встречается у многих авторов. Например, об этом писал известный советский физиолог Н. А. Бернштейн [50]. А одним из первых об отрицательной термодинамической энтропии (термодинамической негэнтропии) в биологии, насколько нам известно, заговорил Э. Шрёдингер в свой широко известной книге «Что такое жизнь?» [51].

Далее Фристон развивает свою мысль следующим образом: «*...набор физиологических и сенсорных состояний, в которых может находиться организм, ограничен, и эти состояния определяют фенотип организма. Математически это означает, что вероятность этих (интероцептивных и экстероцептивных) сенсорных состояний должна иметь низкую энтропию; другими словами, существует высокая вероятность*

того, что система будет находиться в любом из небольшого числа состояний, и низкая вероятность того, что она будет в остальных состояниях. Энтропия — это также средняя собственная информация или «сюрприз»...» [1].

Получается, что, во-первых, здесь Фристон присоединяет сенсорные (то есть *информационные*) состояния к физиологическим состояниям, связанным с гомеостазом, а значит с *термодинамической* энтропией. И почему-то (не вполне понятно почему) сенсорные состояния должны быть связаны с термодинамической негэнтропией. А во-вторых, начав говорить о *термодинамической* негэнтропии, Фристон без приведения достаточных, на наш взгляд, обоснований переходит к *информационной* энтропии сенсорных состояний. С чем связан такой переход в рассуждениях — неясно. Иначе говоря, у Фристона информационная энтропия сенсорных состояний ещё оказывается как-то связанной с термодинамической негэнтропией.

Одно дело, на наш взгляд, когда термодинамическую энтропию можно вывести, аналогично тому, как это делал Больцман [52], из вероятностей состояний ансамбля частиц, где под состояниями понимаются скорости или энергии частиц ансамбля. И совсем другое дело, когда информационная энтропия сенсорных состояний связывается с термодинамической негэнтропией. Это, на наш взгляд, отнюдь не обратная операция по отношению к тому, что делал Больцман. Больцман не связывал состояния системы (ансамбля частиц) ни с какой сенсорикой, а связывал их с вполне определенными физическими понятиями и параметрами — скоростями и энергиями частиц в ансамбле.

Логика такого перехода в рассуждениях из термодинамики в теорию информации вызывает вопросы к Фристону и, как нам представляется, требует подробных разъяснений со стороны автора теории.

У нас нет возражений по поводу негэнтропии живых систем. Однако мы считаем важным различать понятия термодинамической энтропии и информационной энтропии, и в том числе — в приложении к живым системам. При сходных чертах эти понятия всё же не являются тождественными, и этот факт не следует игнорировать.

Таким образом, исходя из рассуждений Фристона, можно сказать, что степени свободы варьирования (в дальнейшем, просто *степени свободы*) своих состояний, которыми может обладать организм и не погибать или не повреждаться, находятся в определённых рамках. Поэтому, для того чтобы поддерживать свои физиологические параметры в пределах этих границ, организму необходимо соблюдать определённый оптимум степеней свободы. Для этого он должен *уменьшать* количество степеней свободы своих состояний.

Аналогичные рассуждения К. Фристон применяет не только ко всему организму, но и к мозгу [1], [5], [11]. В этом случае в роли внешней среды по отношению к мозгу выступает как окружающая организм среда (то есть весь окружающий мир), так и внутренняя среда организма [1], [5]. Внешние по отношению к мозгу состояния среды производят в нервной системе «сенсорные состояния» путём воздействия на сенсорные рецепторы. Другими словами, в этом случае мы имеем систему, состоящую из внешней среды, рецепторов органов чувств (экстероцепторов) и рецепторов внутренних органов (интероцепторов) и мозга. Сенсорные состояния связаны с активностью рецепторов органов чувств и внутренних органов. Эти состояния необходимо держать в определенных границах, совместимых с границами выживания организма.

В теории информации мерой неопределенности состояний системы является величина, называемая *информационной энтропией* [53]-[55]. Устранение неопределённости своих сенсорных состояний, исходя из вышеописанного, согласно Фристону [1]-[12], эквивалентно минимизации информационной энтропии.

По аналогии с энтропией и свободной энергией Гельмгольца [56] из статистической физики (но только по аналогии, не больше!), с информационной энтропией Фристоном [1]-[12] связывается ещё одна величина, называемая им «свободной энергией». Мы пока не будем давать этой величине чёткого определения. В дальнейшем мы сначала уточним это понятие, а в затем раскроем его смысл более подробно. Соответственно, в таком случае минимизация информационной энтропии приводит к минимизации «свободной энергии».

В теории информации и машинном обучении термины «свободная энергия» или «вариационная свободная энергия» одним из первых начал использовать, насколько нам известно, Дж. Хинтон [29], [57]-[60].

Важно принять во внимание то, что понятие «свободная энергия», используемое Хинтоном, а в дальнейшем Фристоном, имеет лишь формальное поверхностное сходство с термодинамической свободной энергией Гельмгольца. Дело в том, что математические формулы для термодинамической свободной энергии Гельмгольца и свободной энергии, используемой Хинтоном и Фристоном, очень похожи. Скорее всего, именно на основании этой аналогии Хинтон ввел данный термин в теорию информации. Фристон в своих работах чаще всего употребляет термин «свободная энергия» без дополнительных прилагательных, хотя термин «вариационная свободная энергия» у него также встречается.

В контексте вышеизложенного, «свободную энергию» Хинтона и Фристона мы будем называть *информационной свободной энергией (ИСЭ)*. Нам кажется, что термин «информационная свободная энергия» более информативен и понятен для восприятия, особенно для неспеци-

алистов в байесовской теории вероятностей. Кроме того, этот термин, как нам кажется, поможет избежать неверных аналогий с соответствующей величиной из термодинамики, так как подчеркивает теоретико-информационный, а не термодинамический характер этой величины. Другими словами, ИСЭ — это не то же самое, что термодинамическая свободная энергия (свободная энергия Гельмгольца) и *ставить знак равенства между этими двумя энергиями категорически неверно*. Об этом неоднократно пишет и сам Фристон [1], [4], [43]. ИСЭ выражается через *информационную энтропию*, а термодинамическая свободная энергия — через *термодинамическую энтропию*. Исходя из этого, не стоит связывать с информационной энтропией такие понятия, ассоциируемые с термодинамической энтропией, как порядок и хаос.

Подчеркнём особенности ИСЭ.

ИСЭ не измеряется ни в джоулях, ни в каких бы то ни было других величинах для измерения энергии из физики. Она не может ни запасаться, ни храниться, ни передаваться, ни преобразовываться, ни тратиться. ИСЭ, по нашему мнению, вполне можно было бы измерять в натах (нат — единица измерения количества информации, определяемая через натуральный логарифм, в отличие, например от бита, определяемого через логарифм по основанию 2 и равная 1.443 бит [61]). Однако, Фристон в своих статьях не упоминает, что ИСЭ может быть измерена в каких-то величинах из теории информации.

В байесовской теории вероятностей, а точнее, в вариационных байесовских методах, очень часто для процедуры вывода, аналогичной той, которая используется в теории Фристона, используется величина, называемая по-разному: «вариационная нижняя оценка», «нижняя оценка обоснованности» или «нижняя оценка предельного правдоподобия» [55]. ИСЭ является величиной, *противоположной* вариационной нижней оценке (нижней оценке обоснованности, нижней оценке предельного правдоподобия), то есть ИСЭ равна вариационной нижней оценке, взятой со знаком минус [55]. В работах Фристона [6], [49] встречается и другой подход для обоснования принципа свободной энергии, состоящий из двух основных идей.

Согласно первой идее предполагается, что любые биологические системы (в том числе и мозг) могут быть описаны как эргодические случайные динамические системы (аттракторы). А гомеостаз выступает как своего рода неотъемлемый атрибут такой системы, связанный с пребыванием состояний этой системы в некотором аттрактивном множестве состояний. Помимо этого, такой эргодический случайный динамический аттрактор имеет, по определению, ещё один атрибут, называемой эргодической плотностью. Эргодическая плотность — это инвариантная вероятностная мера, пропорциональная, в данном случае, количеству времени

в течение которого организм занимает каждое состояние. Или, другими словами, эргодическую плотность можно рассматривать как вероятность нахождения целевой системы в любом из состояний в любой момент времени. А уже в свою очередь с эргодической плотностью можно связать ещё одну величину, называемую энтропией эргодической плотности.

Энтропия эргодической плотности — инвариантная мера, которая суммирует количество различных состояний, которые занимает система.

Согласно второй идее, понятия, связанные эргодической случайной динамической системой, трактуются в байесовском и теоретико-информационном смыслах. В этом случае эргодическая плотность трактуется как предельное правдоподобие, а энтропия эргодической плотности трактуется как информационная энтропия (если говорить более точно — то как дифференциальная энтропия). В результате производится своего рода синтез эргодических случайных динамических систем с теорией информации и байесовской статистикой.

Другими словами, способность живых организмов поддерживать гомеостаз, не повреждаться и не погибать получает при таком подходе [6], [49] теоретико-информационное и статистическое обоснование. Однако и этот подход Фристона не свободен от серьезной критики. Так, в работе [62] подробно разбираются и оспариваются основания считать биологические системы эргодическими случайными динамическими аттракторами. А в работе [63] демонстрируется математическая необоснованность трактовки эргодических случайных динамических систем в теоретико-информационном и байесовском смыслах и прочие нестыковки в теории Фристона с точки зрения математики. Тем не менее в своих последних работах Фристон [64], [65] пытается укрепить математические основания своей теории. Итак, познакомившись с основаниями принципа свободной энергии, можно перейти, наконец, к формулировке этого принципа.

Согласно Фристону, принцип свободной энергии заключается в том, что «*любая самоорганизующаяся система, находящаяся в равновесии с окружающей средой, должна минимизировать свою свободную энергию, противостоя естественной тенденции к беспорядку*» [1].

Это определение нам представляется не очень удачным, потому что в нём не ясно, что значит «противостоя тенденция к беспорядку», и о какой именно свободной энергии идет речь: термодинамической или информационной.

В одной из работ Фристона есть и такое, не проясняющее ситуацию, а скорее ещё более запутывающее, определение: «*Пусть* $m = (R^d, \varphi)$ — *эргодическая случайная динамическая система с пространством состояний* $X = R \times S \in R^d$. *Если внутренние состояния* $r \in R$ *минимизируют свободную энергию, то система подчиняется принципу наименьшего действия и является активной системой.*»

Учитывая эти определения, мы предлагаем сформулировать рассмотренный принцип, в приложении к мозгу, следующим образом:

«Мозг должен минимизировать свою информационную свободную энергию, противостоя возрастанию своей информационной энтропии».

Формулировка принципа свободной энергии именно в таком виде, в котором мы предлагаем, а не в том виде, который предложил Фристон, важна ещё и по следующей причине. Дело в том, что в работах Фристона [1], [5] присутствует некоторый терминологический и смысловой переход (или даже перескок) в рассуждениях: с термодинамики — в теории информации, когда речь идёт о мотивировке Фристоном его принципа свободной энергии. Об этом неожиданном переходе в рассуждениях мы написали выше и он, как нам кажется, не имеет достаточных оснований. Поэтому, читая формулировку принципа свободной энергии, предложенную Фристоном, можно подумать, что речь в ней идёт о термодинамике, но оказывается, нет — речь о теории информации. Наше определение, как мы думаем, не провоцирует таких неоднозначных интерпретаций.

Помимо принципа свободной энергии Фристон использует в своей теории [1], [7], [49] *принцип наименьшего действия*, широко применяемый в физике [56], [66], [67]. Соответственно, в теории Фристона фигурирует ещё одна величина, которую, в связи с соображениями, изложенными выше в отношении ИСЭ, правильнее называть не просто «действием» или «свободным действием», а «*информационным свободным действием*». Математическое выражение, соответствующее принципу наименьшего действия, встречается у Фристона [1], [7], и оно аналогично тому же принципу из физики [56]:

$$S = \int F dt, \tag{1}$$

где S — информационное свободное действие, F — информационная свободная энергия.

Кроме того, под знаком интеграла для аналогичного математического выражения для принципа наименьшего действия из физики стоит лагранжиан. Также и у Фристона [68] свободная энергия — это лагранжиан, то есть некоторая функция от обобщённых координат. Соответственно, для того, чтобы ИСЭ соответствовала форме лагранжиана, Фристоном вводятся в его теорию обобщённые координаты, а ИСЭ, таким образом, является функцией от обобщённых координат.

Фристон называет *информационным свободным действием интеграл по траектории (или пути эволюции) информационной свободной энергии* [3], [4]. А в одной из работ Фристона [49] присутствует более развернутая формулировка этого же: «*внутренние состояния активной системы минимизируют неожиданность..., так что вариация δS действия S по отношению к её внутренним состояниям... обращается в*

нуль». Однако Фристон не приводит подробных пояснений в отношении того, что обозначают эти определения.

Мы, в свою очередь, будем называть информационным свободным действием число, которое получается при интегрировании информационной свободной энергии по времени от момента времени t_1 до момента времени t_2. А принцип наименьшего действия в этом случае заключается в том, что это число будет наименьшим при подстановке в уравнение 1 кривой, соответствующей истинному пути эволюции информационной свободной энергии.

Говоря о стремлении мозга минимизировать неопределённость сенсорных состояний, например, состояний рецепторов органов чувств, Фристон выдвигает важное для его теории положение о том, что мозг не может изменять их *непосредственно*, однако может это делать *опосредованно* через двигательную активность, активно влияя, таким путём, на среду [1], [4], [6]. Поведенческая активность организма изменяет состояние окружающей его среды, а изменение состояний среды, в свою очередь, приводит к изменению сенсорных состояний рецепторов органов чувств.

Однако данное положение о недоступности для мозга прямого управления состояниями сенсорных рецепторов, по нашему мнению, может быть оспорено. Уже довольно давно известно о существовании сенсорных эфферентов, таких как гамма-эфференты [69], оливокохлеарные эфференты [70], ретинопетальные эфференты [71]. По оливокохлеарным и ретинопетальным эфферентам нервные импульсы идут напрямую из головного мозга к соответствующим рецепторам. Благодаря этому головной мозг напрямую управляет чувствительностью сенсорных рецепторов и, следовательно, сенсорными состояниями [70], [71]. Однако, например, гамма-эфферентам Фристон [72] отводит роль своего рода регулятора шума проприоцептивных состояний. В работах Фристона нами не было найдено подробного обсуждения роли оливокохлеарных и ретинопетальных эфферентов. В связи с этим мы считаем недостаточно проработанным вопрос о роли сенсорных эфферентов в теории Фристона.

Следуя логике Фристона, можно сказать, что сенсорные состояния (то есть состояния рецепторов органов чувств), являясь отражением состояний среды, влияют на состояния мозга. Для того, чтобы иметь возможность успевать реагировать на воздействие среды или изменения в среде, мозгу нужно уметь заранее прогнозировать эти воздействия или изменения. Но составлять подобные прогнозы можно лишь на основе той сенсорной информации, которую мозг уже получил ранее.

Главной проблемой, порождающей необходимость в составлении прогнозов, является то, что для организма наиболее выгодно формировать ответную реакцию на изменение среды заранее, ещё *до наступления* это-

го события, до получения сенсорной информации о нём. А это можно сделать только на основании прогноза, составленного на будущее. Можно сказать, что мозг вынужден «работать на опережение» для того чтобы выжить в условиях постоянно меняющейся среды. В дальнейшем мы рассмотрим то, как мозг прогнозирует состояния среды на основе имеющейся сенсорной информации с позиций теории К. Фристона. Важно, что на основе этого прогноза о состояниях среды прогнозируются и будущие предполагаемые сенсорные состояния.

Подводя итог вышеизложенному, можно сказать, что принцип (информационной) свободной энергии К. Фристона имеет потенциальный объяснительный потенциал. Однако этот принцип игнорирует возможность прямого управляющего влияния мозга на сенсорные состояния и допускает неясности в отношении таких понятий как «термодинамическая энтропия», с одной стороны, и «информационная энтропия», с другой стороны. Кроме того, как мы уже упоминали, остаётся неясным, что же подразумевается под термином «состояние». Эти моменты, как нам представляется, являются существенными недоработками теории Фристона.

3. Фристоновская формулировка гипотезы байесовского мозга (байесовского кодирования)

В этом разделе мы продолжаем рассматривать систему, состоящую, с одной стороны, из среды, экстерорецепторов органов чувств и интерорецепторов внутренних органов, которые мы называем *сенсорными рецепторами*, и, с другой стороны, мозга, не вдаваясь в конкретные детали его устройства. Работу мозга с информацией мы будем рассматривать с позиций байесовской теории вероятностей, предоставляющей математический аппарат для такого рассмотрения. В дальнейшем, при таком подходе, речь идёт скорее не о работе реального мозга, а о работе абстрактного, «байесовского мозга». А для того чтобы рассматривать работу мозга таким образом, сначала мы введём основные понятия, необходимые для такого рассмотрения.

Как мы писали в предыдущем разделе, внешние по отношению к мозгу состояния среды производят сенсорные состояния, связанные с активностью сенсорных рецепторов. Иначе говоря, перефразируя Фристона [1], [2], [11], [43], можно сказать, что *мозгом воспринимаются не сами объекты внешней среды непосредственно, а сенсорные состояния, которые являются результатом действия внешней среды на сенсорные рецепторы*. О похожих по смыслу вещах писал ещё Гельмгольц [13], [15] (о подобных вещах помимо Гельмгольца писали ещё Платон с его пред-

ставлением о «ноуменах» [73] и И. Кант — как о «вещах в себе» [74]). Другими словами, состояния среды являются для самого мозга *ненаблюдаемыми*, не доступными непосредственно величинами, а сенсорные состояния являются *наблюдаемыми*, доступными непосредственно самим мозгом величинами. И, соответственно, *состояния внешней среды «доступны» для мозга только опосредованно*, через сенсорные состояния (наблюдаемые мозгом величины), являющиеся зависимыми от состояний внешней среды (ненаблюдаемых состояний), так как сенсорные состояния порождаются состояниями среды [1], [2], [11], [43].

По этому поводу можно образно сказать, что между внешней средой и мозгом есть своего рода «стена» за которую мозг не может выйти и которую не может преодолеть. А сенсорные рецепторы в этом случае выполняют роль своего рода «переводчика», который переводит информационные сигналы с «языка окружающей среды» на «язык мозга». Среда непосредственно влияет только на состояние «переводчика», воспринимающего её «язык», но не на сам мозг. Без такого «рецепторно-сенсорного перевода» информационные сигналы окружающей среды для мозга просто недоступны. Поскольку состояния среды являются ненаблюдаемыми величинами, то о том, в каком состоянии находится внешняя среда, можно говорить только в предположительном, *вероятностном* ключе, выдвигая *гипотезы* о ней [1], [2], [11], [43]. О близких по смыслу вещах, правда на другую тему, пишет, например, известный физик-теоретик Ричард Фейнман: «*Откуда мы всё–таки знаем, что атомы существуют? А здесь идёт в ход уже описанный приём: мы предполагаем их существование, и все результаты один за другим оказываются такими, как мы предскажем, какими они должны быть, если всё состоит из атомов*» [75].

Такая *вероятность, которая позволяет работать с ненаблюдаемыми величинами вне зависимости от природы процесса, которому подчиняется их поведение, называется байесовской вероятностью* [76]-[78].

В соответствии с этим подходом, мозг решает так называемую *обратную задачу* [1], [2], [11], [43]. Он по *результатам* (в роли результата выступают сенсорные состояния) делает суждение о *причинах* (в роли причин выступают состояния среды), приведших к этому результату. Обратной эта задача называется потому, что она является противоположной по отношению к *прямой задаче*, то есть задаче, где мы на основе исходных причин делаем предположения о возможных следствиях (результатах).

Другими словами, обобщая написанное Фристоном в его работах [1], [4], [5], можно сказать, что идея о том, что мозг делает суждения о состояниях среды по сенсорным состояниям, аналогична по смыслу идее «бессознательного вывода», предложенной Гельмгольцем.

Интересно отметить, что прямая задача — это по сути вывод следствий в классической и булевой логиках. Однако эти логики не способны решать обратные задачи, в то время как байесовский подход в теории вероятностей позволяет решать как прямые задачи, так и обратные, таким образом являясь обобщённием правил вывода классической и булевой логик [70], [71], [76]-[78].

Соответственно, предполагая байесовские вероятности состояний внешней среды и принимая во внимание зависимость от них сенсорных состояний, мы можем говорить, в свою очередь, о *байесовской вероятности сенсорных состояний*, то есть о вероятности наступления сенсорных состояний при условии, что внешняя среда имеет некоторое состояние [1], [2], [4], [5], [11], [43]. Такая вероятность называется *условной вероятностью или условной байесовской вероятностью*. Такой подход соответствует байесовскому подходу в теории вероятностей, который позволяет работать как с ненаблюдаемыми (скрытыми, латентными) величинами, так и с наблюдаемыми величинами [76]-[78].

Для точного подсчёта как условной вероятности, так любых других байесовских вероятностей используется, как известно, теорема Байеса [53], [54], [79]-[81].

Теорема Байеса (или формула Байеса) в теории Фристона является той первоосновой, которая пронизывает весь математический аппарат этой теории, и вокруг которой он вращается. Можно даже сказать, что те основные *уравнения теории Фристона*, к которым мы придём через некоторое время — *это изменённая почти до неузнаваемости, преобразованная и дополненная добавочными членами, но, по сути, всё та же формула Байеса. А все вероятности, которые мы используем в этой статье, являются байесовскими вероятностями.*

Согласно теореме Байеса условная вероятность события B при условии, что событие A наступило, равна произведению условной вероятности события A при условии B и безусловной вероятности события A, делённой на безусловную вероятность события B [53]-[55]. Или:

$$p(B|A) = \frac{p(A|B)p(A)}{p(B)}, \tag{2}$$

где $p(A|B)$ — условная вероятность события A при условии, что событие B наступило; $p(B|A)$ — условная вероятность события B при условии A; $p(A)$ — безусловная вероятность события A; $p(B)$ — безусловная вероятность события B.

Буква «p» (от англ. probability) во всех формулах, которые мы используем, обозначает вероятность.

В байесовской теории вероятностей величина $p(A|B)$ называется *правдоподобием*, величина $p(B|A)$ называется *апостериорной вероят-*

ностью, величина $p(A)$ называется *предельным или интегрированным правдоподобием, или обоснованностью, или обоснованностью модели*, и $p(B)$ — *априорной вероятностью* [53]-[55].

В нашем случае событие A — это сенсорные состояния, событие B — состояния среды.

Стоит сразу отметить, что мы в дальнейшем будем иметь дело не с какими-то одиночными вероятностями, соответствующими соотношениям между этими состояниями, а с *вероятностными распределениями*, определяющими эти соотношения. В соответствии с этим будут производится не точечные оценки тех или иных вероятностей, а будут оцениваться вероятностные распределения.

Теорема Байеса принимает тогда следующий вид: вероятностное распределение нахождения внешней среды в некотором состоянии при условии сенсорных состояний равна произведению вероятностного распределения наступления сенсорных состояний при условии, что внешняя среда находится в некотором состоянии, и вероятностного распределения состояний внешней среды, делённой на вероятностное распределение сенсорных состояний. Обозначив $s = s(t)$ — сенсорные состояния, $\upsilon = \upsilon(t)$ — состояния среды (Буква t в s и υ обозначает, что сенсорные состояния и причинные состояния среды изменяются во времени, однако, для экономии места и меньшей нагроможденности формул, мы будем использовать s и υ, не указывая зависимость этих переменных от времени) и заменив на s и B на υ, получим:

$$p(\upsilon|s) = \frac{p(s|\upsilon)p(\upsilon)}{p(s)}, \tag{3}$$

где υ — состояния среды, или ненаблюдаемые переменные, или причины сенсорных данных, которые состоят из скрытых (латентных) состояний x, скрытых причин ν и параметров состояний внешней среды θ и γ; s — сенсорные данные, или наблюдаемые переменные; $p(\upsilon|s)$ — апостериорное распределение вероятностей; $p(s|\upsilon)$ — правдоподобие; $p(s)$ — предельное правдоподобие, интегрированное правдоподобие, обоснованность, или нормировочная константа; $p(\upsilon)$ — априорное распределение вероятностей состояний среды (скрытых, или латентных, состояний x и параметров состояний внешней среды θ) [1], [2], [4], [5], [11], [43].

Теперь давайте разберёмся в этих обозначениях более подробно.

Для начала скажем, что мы употребляем понятия «состояния среды», «причинные состояния среды» или «ненаблюдаемые переменные», или «ненаблюдаемые величины», или «причины сенсорных данных» синонимично и равнозначно. Стоит ещё добавить, что вместо термина «состояния среды» мы в дальнейшем будем чаще употреблять термин «причин-

ные состояния среды» для того, чтобы у читателей не возникали ложные ассоциации с понятием «среда» из экологии.

Сначала разберёмся в отношении того, что такое *причинные состояния среды* υ. Как мы написали выше, υ состоят из скрытых (латентных) состояний x, скрытых причин ν и параметров состояний внешней среды θ и γ. В статьях Фристона есть путаница по этому поводу. Например, в одной из своих работ [11] он пишет следующее: «*Причины — это просто состояния процессов, генерирующих сенсорные данные. Причины могут быть категориальными по своей природе, например, идентичность лица или семантическая категория, к которой принадлежит объект. Другие могут быть параметрическими, например положение объекта...*» и «*...Причины — это количества или состояния, которые необходимы для определения продуктов процесса, генерирующего сенсорную информацию...*». И называет в качестве примера причин такие вещи как скорость конкретного объекта, направление световых лучей и т. д. [11]. Причём из текста работы не совсем ясно о чём же идёт речь: о скрытых состояниях, скрытых причинах, параметрах или обо всём этом сразу. В другой работе [82] Фристон, первоначально вводя термин «скрытые состояния», затем пишет, что скрытые состояния делятся на скрытые динамические состояния и скрытые причины, а затем по ходу текста термин «скрытые состояния» вновь употребляется без каких-либо дополнительных пояснений о том, что речь идет о скрытых динамических состояниях. Другими словами, из этой работы [82], а также из других работ [1], [3]-[5], [7] можно подумать, что скрытые состояния делятся на скрытые состояния и скрытые причины. А ещё в одной из работ [123] Фристон исходно использует термин «причины» для обозначения величины, аналогичной нашей υ. А затем в этой же работе выясняется, что причины υ делятся на скрытые состояния x и причины ν. Причем причины ν, как пишет Фристон, являются состояниями, как и, собственно, скрытые состояния x. Иначе говоря, исходя из этой работы [123], можно подумать, что причины делятся на скрытые состояния и причины. Общим знаменателем этих работ, тем не менее, является то, что есть некоторое общее понятие, которое делится на две категории: скрытые состояния и причины. Нужно сказать, что и наше определение для этого общего понятия, такое как «состояния среды» или «причинные состояния среды», тоже может показаться несколько неудачным. Однако достаточно сложно предложить термин, который был бы близок к тому, что употребляет Фристон, и был бы свободен от вышеописанных недостатков. Мы считаем, что эти наши пояснения освободят читателя от возможного недопонимания работ Фристона в отношении скрытых состояний и причин.

Исходя из написанного выше, мы, для состояний среды или «причин сенсорных данных» υ, для начала предлагаем такое определение: *причинные состояния среды υ — это любые состояния окружающей среды, функции от этих состояний и функции, связанные с этими состояниями, информацию о которых мозг получает через сенсорные рецепторы, то есть через изменения сенсорных состояний.*

Для скрытых состояний и скрытых причин Фристон даёт следующие определения. Скрытые причины ν — это функции от скрытых состояний [82]. «*Состояния ν называются причинами, входами или источниками «могут быть детерминированными, стохастическими или и теми, и другими»* [123]. Состояния x называются скрытыми состояниями, «*потому что их редко можно наблюдать напрямую*» и они «*опосредуют влияние ввода на вывод и наделяют систему памятью*» [123]. В следующей главе эти определения будут нами расширены и уточнены.

Заметим, что байесовский подход в теории вероятностей допускает работать с ненаблюдаемыми (скрытыми, латентными, неопределёнными) состояниями точно так же, как с обычными случайными величинами, вне зависимости от природы процесса их породившего (здесь имеется ввиду, что процесс, порождающий ненаблюдаемые состояния может быть как случайным, так и детерминистическим). В этом случае в байесовской интерпретации вероятность связывается с уровнем доверия (belief). И с помощью теоремы Байеса в таком случае оценивается доверие предположению (гипотезе) о распределении ненаблюдаемой (скрытой, латентной) величины (причины) до и после принятия во внимание сенсорных данных [53]-[55].

Для сенсорных состояний s Фристон чёткого определения не приводит, поэтому мы дадим наше собственное определение. Под *сенсорными состояниями s* мы понимаем *состояния сенсорных рецепторов*, которые являются зависимыми от среды. Зависимость сенсорных состояний от причин заключается в том, что причины порождают сенсорные данные. Процесс порождения причинами сенсорных данных называется *порождающим процессом.* «Задать порождающий процесс» — означает задать математическую модель того, как причины порождают сенсорные данные. Сенсорные состояния являются наблюдаемыми величинами. В роли «наблюдателя» сенсорных состояний выступает мозг.

Апостериорное распределение вероятностей (или просто апостериорное распределение) Фристон определяет так [[1]]: «*апостериорное распределение вероятностей — это распределение вероятностей состояний среды или параметров модели, учитывая некоторые данные, то есть вероятностное отображение (отображение — математический термин) из наблюдаемых данных в состояния среды*».

Мы, в свою очередь, называем апостериорным распределением вероятностей условное вероятностное распределение нахождения внешней среды в некотором состоянии при учёте данных сенсорных состояний. Апостериорное вероятностное распределение $p(\upsilon|s)$ можно также назвать *вероятностным распределением причинных состояний среды* υ при условии, что сенсорные данные s получены. Процедура оценки или вычисления апостериорной вероятности или апостериорного вероятностного распределения называется *байесовским выводом* [53]-[55].

Априорное распределение вероятностей (или просто априорное распределение), по Фристону [1], — это вероятностное распределение состояний среды, которое кодирует убеждения (beliefs) об этих состояниях *перед* наблюдением сенсорных данных. Мы называем априорным распределением вероятностей (априорным распределением вероятностей причинных состояний среды) $p(\upsilon)$ величину, которая выражает *предварительные предположения* о причинных состояниях среды до учёта реальных сенсорных данных. Иначе это можно назвать *гипотезой о ненаблюдаемых состояниях перед непосредственным опытом*, то есть перед получением сенсорных данных, верифицирующих (подтверждающих или отвергающих) данную гипотезу.

Правдоподобие $p(s|\upsilon)$, согласно Фристону [84], определяет то, каким образом состояния среды приводят к сенсорным данным. Правдоподобием мы называем условное распределение вероятностей наступления сенсорных состояний при условии, что внешняя среда находится в некотором причинном состоянии. Проще можно сказать, что *правдоподобие показывает, каким образом наступление сенсорных состояний зависит от причинных состояний среды.*

Предельное правдоподобие (обоснованность) $p(s)$ по Фристону — это вероятностное распределение выборки некоторых данных в рамках конкретной модели [85].

Мы называем предельным правдоподобием вероятностное распределение сенсорных состояний, при котором причинные состояния среды υ исключены с помощью операции маргинализации, то есть интегрирования произведения правдоподобия $p(s|\upsilon)$ и априорного распределения $p(\upsilon)$ по причинным состояниям среды υ.

Помимо вышеописанных величин, в теории Фристона [1], [2], [11], [43] фигурирует ещё одна величина, которая может быть получена преобразованием числителя в правой части уравнения 3 следующим образом:

$$p(\upsilon|s) = \frac{p(s,\upsilon)}{p(s)}, \tag{4}$$

эта новая величина $p(s,\upsilon)$ — совместное распределение вероятностей.

Совместным распределением вероятностей $p(s, \upsilon)$ называется величина равная произведению правдоподобия $p(s|\upsilon)$ на априорную вероятность $p(\upsilon)$ [86]. Совместное распределение вероятностей в байесовской теории вероятностей называется *порождающей моделью*. Фристон приводит такое определение порождающей модели: *«порождающая модель — это вероятностная модель зависимостей между причинами и следствиями (данными), из которых могут быть получены выборки»* [1].

Мы, со своей стороны, можем дать такое определение порождающей модели: порождающая (генеративная) модель — это вероятностная (статистическая) модель совместного распределения вероятностей причинных состояний внешней среды (причин сенсорных данных) и сенсорных данных (для точного понимания и корректного употребления термина «порождающая модель» его всегда можно заменить на словосочетание «совместное распределение вероятностей» или «произведение правдоподобия на априорное распределение вероятностей»).

А статистическая модель — это, в свою очередь, набор математических функций, которые описывают поведение объектов в терминах случайных величин и связанных с ними распределений вероятностей [86]. (Заметим, что «статистическая модель» является более общим понятием, чем «вероятностная модель». Иначе говоря, статистическая модель представляет из себя набор вероятностных моделей [87]. Хотя с точки зрения теории вероятностей не совсем корректно эти понятия использовать синонимично, однако, в контексте нашей статьи, этот момент не является принципиальным.) Порождающая модель называется порождающей потому, что мы можем сделать выборку из совместного распределения для получения новых данных вместе с их метками [81].

Задать вероятностную (статистическую) модель — это значит задать совместное распределение вероятностей [53]-[55]:

$$p(s, \upsilon) = p(s|\upsilon)p(\upsilon). \tag{5}$$

(При использовании теории Фристона в собственных исследованиях для задания порождающих моделей можно использовать графовое представление порождающей модели. Иначе можно сказать, что порождающую модель можно представить как ориентированный граф. Такое представление порождающей модели может быть полезно для того, чтобы не забыть условные вероятности между одними узлами и другими узлами этой сети (графа) и правильно задать порождающую модель.) А сам процесс, при котором выборочные информационные данные (сенсорные сэмплы) используются для получения некоторого представления о состояниях среды, в теории Фристона называется статистическим выводом [1]. Теперь, когда мы охарактеризовали основные понятия, фигурирующие

в фристоновской формулировке гипотезы байесовского мозга, можно перейти к определению для этой гипотезы.

Фристон предлагает такое определение: мозг использует внутренние вероятностные (порождающие) модели для обновления апостериорных вероятностей, используя сенсорную информацию, в (приблизительно) оптимальном, по Байесу, виде [1].

В литературе по байесовской теории вероятностей нам удалось найти только одну формулировку того, что такое оптимальность по Байесу: вероятностная модель является оптимальной по Байесу, если выполняются предположения об истинном апостериорном распределении [81].

Во введении мы уже приводили гипотезу байесовского кодирования (байесовского мозга), однако, по нашему мнению, и фристоновская формулировка гипотезы байесовского мозга, и та, которая была приведена нами во введении, недостаточно удачны, поэтому мы приведём здесь свой вариант гипотезы байесовского мозга. Гипотеза байесовского мозга состоит в следующем:

- Мозг представляет сенсорную информацию в форме вероятностных распределений: как совместного распределения вероятностей сенсорных данных и причинных состояний среды (причин сенсорных данных), так и предельного правдоподобия (обоснованности) сенсорных данных.

- Мозг кодирует и вычисляет оптимальным, по Байесу, образом функции плотности вероятности апостериорного распределения причинных состояний среды и сенсорных данных или приближения к функциям плотности вероятности апостериорного распределения (вариационное распределение состояний среды или плотность распознавания). Здесь встречается термин «функция плотности вероятности» или просто «плотность вероятности», которая является ничем иным как способом задания вероятностного распределения. Для дискретной случайной величины вероятностное распределение задается как функция вероятности, а для непрерывной случайной величины — как плотность вероятности. Так как плотность вероятности является способом задания вероятностного распределения непрерывной случайной величины и учитывая то, что в нашей статье мы рассматриваем теорию Фристона в непрерывной формулировке, то в дальнейшем мы будем использовать термины «вероятностное распределение» и «плотность вероятности» как взаимозаменяемые. О том, что такое «вариационное распределение причинных состояний среды (плотность распознавания)» — будет написано ниже.

Из гипотезы байесовского мозга и из рассуждений, представленных выше, по нашему мнению, вытекает ряд следствий, нигде явно не упоминаемых Фристоном:

1) Мозг строит вероятностные (статистические) модели окружающей среды. У Фристона речь идёт не о каких-то там «нервных моделях», а о моделях именно вероятностных (статистических).

2) Вероятностные распределения, используемые мозгом, имеют марковское свойство.

3) Работа мозга может быть описана в согласии с теоремой Байеса.

Теперь перейдем к математическим преобразованиям, связывающим теорему Байеса и, соответственно, гипотезу байесовского мозга, с одной стороны, и ИСЭ, введенную в предыдущем разделе, с другой стороны. Заметим, что основную часть этих преобразований Фристон не приводит в своих работах, ограничиваясь кратким изложением [1], [2], [11], [43], однако в работах других авторов встречаются аналогичные, приводимым нами, математические преобразования [45], [46].

Помимо уравнений 3 и 4, описанные выше величины можно представить по-другому, а именно:

$$p(\upsilon|s) = \frac{p(s|\upsilon)p(\upsilon)}{\int p(s,\upsilon)d\upsilon} = \frac{p(s,\upsilon)}{\int p(s,\upsilon)d\upsilon}, \tag{6}$$

Тогда предельное правдоподобие, или обоснованность, можно выразить следующим образом:

$$p(s) = \frac{p(s,\upsilon)}{p(\upsilon|s)}, \tag{7}$$

При работе с плотностями вероятностей удобнее работать с логарифмическими значениями (натуральный логарифм) соответствующих плотностей, поэтому логарифмируя левую и правую части уравнения 7 получим:

$$\ln p(s) = \ln \frac{p(s,\upsilon)}{p(\upsilon|s)}, \tag{8}$$

Расчёт предельного правдоподобия (обоснованности) $p(s)$ сопряжён со значительными, часто непреодолимыми, трудностями, потому что эта величина представляет из себя интеграл: $p(s) = \int p(s,\upsilon)d\upsilon$. Поэтому вместо расчёта этого интеграла, согласно вариационным байесовским методам, вводится новая функция — вариационное распределение $q(\upsilon)$, которая в теории Фристона называется «плотностью распознавания». Соответственно, Фристоном вводится дополнительное предположение: мозг

должен кодировать плотность распознавания $q(\upsilon)$ (несмотря на важность этого предположения для всего последующего формулирования математического аппарата теории Фристона, в изученных нами работах Фристона нигде отдельно это предположение явно не выделяется). Апостериорное распределение набора ненаблюдаемых (скрытых, латентных) причинных состояний среды υ над наблюдаемыми сенсорными данными s в том случае приближается вариационным распределением $q(\upsilon)$ (плотностью распознавания, прокси-функцией) (вариационные байесовские методы (или, что то же самое, «ансамблевое обучение») — это методы вычисления апостериорного распределения вероятностей, основанные на введении вариационного распределения (прокси-функции, плотности распознавания). Или, проще говоря, плотность распознавания $q(\upsilon)$ (вариационное распределение или вариационное распределение причинных состояний среды) становится приблизительным апостериорным распределением $p(\upsilon|s)$:

$$p(\upsilon|s) \approx q(\upsilon). \tag{9}$$

Плотность распознавания (прокси-функция, вариационное распределение) $q(\upsilon)$ нормируется как:

$$\int q(\upsilon)d\upsilon = 1. \tag{10}$$

Учитывая уравнение 10, согласно неравенству Йенсена [53]-[55] уравнение 8 можно переписать так:

$$\ln p(s) = \int q(\upsilon) \ln \frac{p(s,\upsilon)}{p(\upsilon|s)} d\upsilon, \tag{11}$$

или, соблюдая тождество, так:

$$\ln p(s) = \int q(\upsilon) \ln \frac{p(s,\upsilon)q(\upsilon)}{p(\upsilon|s)q(\upsilon)} d\upsilon, \tag{12}$$

а произведение под знаком натурального логарифма в правой части уравнения 12 может быть переписано как сумма натуральных логарифмов:

$$\ln p(s) = \int q(\upsilon) \left(\ln \frac{q(\upsilon)}{p(\upsilon|s)} + \ln \frac{p(s,\upsilon)}{q(\upsilon)} \right) d\upsilon. \tag{13}$$

Интеграл суммы равен сумме интегралов:

$$\ln p(s) = \int q(\upsilon) \ln \frac{q(\upsilon)}{p(\upsilon|s)} d\upsilon + \int q(\upsilon) \ln \frac{p(s,\upsilon)}{q(\upsilon)} d\upsilon. \tag{14}$$

Теперь перепишем 2-й интеграл в правой части уравнения 14, поменяв местами $p(s, \upsilon)$ и $q(\upsilon)$. Это приводит и к смене знака перед этим интегралом на противоположный. В итоге получим:

$$\ln p(s) = \int q(\upsilon) \ln \frac{q(\upsilon)}{p(\upsilon|s)} d\upsilon - \int q(\upsilon) \ln \frac{q(\upsilon)}{p(s, \upsilon)} d\upsilon. \tag{15}$$

или для краткости:

$$\ln p(s) = D[q(\upsilon)||p(\upsilon|s)] - F(s, q(\upsilon)). \tag{16}$$

Первый член в правой части уравнений 15 и 16 — это дивергенция (расхождение) Кульбака-Лейблера $D(q(\upsilon)||p(s|\upsilon))$ между вариационным распределением $q(\upsilon)$ и апостериорным распределением $p(s|\upsilon)$, второй член в этих уравнениях — это информационная свободная энергия $F(s, q(\upsilon))$, то есть:

$$F(s, q(\upsilon)) = \int q(\upsilon) \ln \frac{q(\upsilon)}{p(s, \upsilon)} d\upsilon. \tag{17}$$

Хотя, совершенно равнозначное к данному, это представление информационной свободной энергии может иметь вид:

$$F(s, q(\upsilon)) = -\int q(\upsilon) \ln \frac{p(s, \upsilon)}{q(\upsilon)} d\upsilon. \tag{18}$$

Другими словами, исходя из уравнений 6-17, мы, вслед за Фристоном, устанавливаем связь между теоремой Байеса и, соответственно, гипотезой байесовского мозга, и ИСЭ, введенной в первом разделе. И оказывается, что ИСЭ выражается через вероятностные зависимости между ненаблюдаемыми состояниями и сенсорными данными. Так, если проанализировать уравнение 18, видно, что чем меньше будет разница между порождающей моделью и плотностью распознавания, то тем меньше будет информационная свободная энергия и, следовательно, тем точнее мозг предсказывает изменения среды, причины сенсорного входа. То есть тем ближе вероятностное распределение, производимое плотностью распознавания $q(\upsilon)$, к истинному распределению причинных состояний среды υ. А чем точнее мозг предсказывает причинные состояния среды, тем точнее он предсказывает свой сенсорный вход, так как последний зависим от состояний среды. Чем меньше информационная свободная энергия, тем менее неопределёнными для мозга являются будущие сенсорные состояния и скрытые причинные состояния среды, их породившие.

В этом контексте ***восприятие*** предстает перед нами как ***статистический вывод о причинных состояниях среды (причинах сенсорного входа)***, а не просто как измерение или отбор сенсорных данных [6].

Кроме того, взглянув на уже описанное, мы приходим к ещё одному следствию из гипотезы байесовского мозга в теории Фристона: о том, что ***всю деятельность мозга можно рассматривать как работу с вероятностными репрезентациями ненаблюдаемых и наблюдаемых состояний, носящую характер статистического вывода о ненаблюдаемых состояниях на основе наблюдаемых***.

Заметим, что если посмотреть на второй интеграл в правой части уравнения 15, соответствующий информационной свободной энергии, то видно, что он похож на первый интеграл из той же части, того же уравнения.

Иначе говоря, следуя логике К. Фристона, мы можем сделать собственный вывод о том, что, ***информационная свободная энергия — это своего рода дивергенция Кульбака-Лейблера между вариационным распределением*** $q(\upsilon)$ ***и плотностью совместного распределения вероятностей*** $p(s, \upsilon)$***, то есть порождающей моделью***.

Если вариационное распределение (плотность распознавания) $q(\upsilon)$ совпадает с плотностью совместного распределения вероятностей (порождающей моделью) $p(s, \upsilon)$, то есть $q(\upsilon) = p(s, \upsilon)$, то имеем:

$$\int q(\upsilon) \ln \frac{q(\upsilon)}{p(s, \upsilon)} d\upsilon = \int q(\upsilon) \ln 1 d\upsilon = \int 0 d\upsilon = Const. \tag{19}$$

Как мы видим из уравнения 19, если вариационное распределение $q(\upsilon)$ совпадает с порождающей моделью $p(s, \upsilon)$, то расхождение Кульбака-Лейблера $D(q(\upsilon)||p(s, \upsilon))$ равно константе. Вспомним, однако, что и предельное правдоподобие (обоснованность) - это нормировочная константа. Иначе говоря, если вариационное распределение $q(\upsilon)$ (плотность распознавания) практически совпадает с порождающей моделью $p(s, \upsilon)$, то информационная свободная энергия $F(s, q(\upsilon))$ приближается к нулю. Это можно интерпретировать в том ключе, что ИСЭ приближается к нулю в случае, если мозг спрогнозировал причинные состояния среды с точностью приближающейся к 100

И наоборот, чем сильнее вариационное распределение $q(\upsilon)$ (плотность распознавания) отличается от порождающей модели $p(s, \upsilon)$ тем больше расхождение Кульбака-Лейблера $D(q(\upsilon)||p(s, \upsilon))$ между этими распределениями и, соответственно, тем больше значение, принимаемое информационной свободной энергией F(s, $q(\upsilon)$). Или же: чем менее точно мозг прогнозирует причинные состояния среды, тем выше ИСЭ. Поэтому, чтобы осуществить байесовский вывод, необходимо с помощью оптимизации минимизировать информационную свободную энергию $F(s, q(\upsilon))$.

Другими словами, принимая во внимание формулировку принципа свободной энергии, о том, что мозг должен минимизировать свою ин-

формационную свободную энергию, можно сказать, что осуществление байесовского вывода, в контексте теории Фристона, эквивалентно принципу информационной свободной энергии. Помимо уравнений 17 и 18, Фристон в своих работах [1]-[12] дает ещё несколько представлений для ИСЭ. Несколько изменив уравнение 16, получим ещё одно представление ИСЭ, встречающееся у Фристона:

$$F(s, q(\upsilon)) = D[q(\upsilon)||p(\upsilon|s)] - \ln p(s). \tag{20}$$

Информационная свободная энергия — это расхождение Кульбака-Лейблера $D(q(\upsilon)||p(s,\upsilon))$ между вариационным распределением $q(\upsilon)$ и апостериорным распределением $p(\upsilon|s)$ минус сенсорный сюрприз $lnp(s)$ (логарифмическое предельное правдоподобие, логарифмическая обоснованность, а в дискретной формулировке теории Фристона — собственная информация (собственная информация — это количественная мера неожиданности конкретного исхода [53])) [1].

В своих работах [1], [4], [5] Фристон приводит представление ИСЭ ещё в четырёх вариантах. В первом из этих вариантов определение ИСЭ не связано с конкретной формулой и звучит следующим образом:

Информационная свободная энергия (ИСЭ) — это, согласно Фристону, мера, которая «ограничивает или лимитирует неожиданность при выборке некоторых данных от данной порождающей модели»: «An information theory measure that bounds or limits (by being greater than) the surprise on sampling some data, given a generative model» [1]. Это означает, что неожиданность выборочных данных не может быть больше ИСЭ. Три других представления связаны с конкретными формулами.

Так как $p(s,\upsilon) = p(s|\upsilon)p(\upsilon)$, то в уравнении 17 $p(s,\upsilon)$ может быть заменена на $p(s|\upsilon)p(\upsilon)$, то получаем:

$$\begin{aligned} F(s, q(\upsilon)) &= \int q(\upsilon) \ln \frac{q(\upsilon)}{p(s,\upsilon)} d\upsilon \\ &= \int q(\upsilon) \ln \frac{q(\upsilon)}{p(s|\upsilon)p(\upsilon)} d\upsilon \\ &= \int q(\upsilon) \ln \frac{q(\upsilon)}{p(\upsilon)} d\upsilon - \int q(\upsilon) \ln p(s|\upsilon) d\upsilon \\ &= D[q(\upsilon)||p(\upsilon)] - \langle \ln p(s|\upsilon) \rangle_{q(\upsilon)}. \end{aligned} \tag{21}$$

Другими словами, ***ИСЭ*** — это $D(q(\upsilon)||p(\upsilon))$, то есть дивергенция Кульба-ка-Лейблера между вариационным распределением $q(\upsilon)$ и априорным распределением причинных состояний среды $p(\upsilon)$ [4]) минус «accuracy» (адекватный перевод термина «accuracy» на русский язык непрост и вызывает проблемы в различных областях науки и техники [88], поэтому мы воздержимся от неточных вариантов перевода этого

термина русский язык) $\langle \ln p(s|\upsilon) \rangle_{q(\upsilon)}$. «Accuracy» показывает, насколько вероятны сенсорные данные при конкретных исходах причинных состояний среды. Кроме того, можно ввести величину равную - $\langle \ln p(s|\upsilon) \rangle_{q(\upsilon)}$, которую можно назвать перекрёстной информационной энтропией между вариационным распределением $q(\upsilon)$ и правдоподобием $p(s|\upsilon)$, то есть:

$$H(q(\upsilon), p(s|\upsilon)) = -\int q(\upsilon) \ln p(s|\upsilon) d\upsilon. \tag{22}$$

Уравнение 17 можно преобразовать иначе:

$$\begin{aligned} F(s, q(\upsilon)) &= -\int q(\upsilon) \ln p(s, \upsilon) d\upsilon + \int q(\upsilon) \ln q(\upsilon) d\upsilon \\ &= L(s, \upsilon) - H(q(\upsilon)), \end{aligned} \tag{23}$$

где $L(s, \upsilon) = -\int q(\upsilon) \ln p(s, \upsilon) d\upsilon$ — информационная энергия Гиббса, $H(q(\upsilon)) = -\int q(\upsilon) \ln q(\upsilon) d\upsilon$ — дифференциальная энтропия вариационного распределения.

Иначе говоря, **ИСЭ** — это информационная энергия Гиббса $L(s, \upsilon)$ (ожидаемая энергия, или средняя энергия) плотности совместного распределения вероятностей $p(s, \upsilon)$ минус дифференциальная энтропия вариационного распределения $H[q(\upsilon)]$.

Фристон в своих работах [7] приводит такое определение для информационной энергии Гиббса $L(s, \upsilon)$: ***информационная энергия Гиббса*** *— это логарифмическая вероятность, которая сообщает о совместной неожиданности сенсорных данных и вызывающих их причинных состояний среды.*

Однако, внимательно приглядевшись к представленным уравнениям, мы можем дать другое определение. Как нетрудно заметить, посмотрев на уравнения 21 и 23, информационная энергия Гиббса $L(s, \upsilon)$ — это величина обратная accuracy $\langle \ln p(s|\upsilon) \rangle_{q(\upsilon)}$. Или, сравнив уравнения 22 и 23, мы обнаруживаем, что *информационная энергия Гиббса $L(s, \upsilon)$ — это величина равная перекрёстной информационной энтропии между вариационным распределением $q(\upsilon)$ и правдоподобием $p(s|\upsilon)$.*

Для дифференциальной энтропии вариационного распределения $H[q(\upsilon)]$ у Фристона нам не удалось найти чёткого определения, поэтому мы рискнем дать собственное определение: *дифференциальная энтропия вариационного распределения $H[q(\upsilon)]$ — это мера неопределённости вариационного распределения $q(\upsilon)$.*

Из этих пяти формулировок, именно представление ИСЭ, используемое в уравнении 23, и использует в своих работах Фристон для более детального формулирования математического аппарата своей теории [2], [3], [7].

Прежде чем подводить итог описанного в этой главе, необходимо сделать ещё две вещи.

Во-первых, следует вспомнить то, о чём мы писали в предыдущей главе, и привести то, что описано здесь, в соответствие с тем, что было описано ранее. Как было сказано в предыдущей главе, для того чтобы ИСЭ соответствовала форме лагранжиана, Фристоном [7] вводятся в его теорию обобщённые координаты, а ИСЭ, таким образом, является функцией (а ещё точнее функционалом) от обобщённых координат. Другими словами, так как ИСЭ является функционалом от сенсорных состояний s и ненаблюдаемых состояний υ, то это означает, что эти состояния должны быть представлены в форме обобщённых координат. Обобщённые координаты в теории Фристона представляют из себя производные по времени от функций, описывающих состояния [2]-[5]. Эти производные представляют из себя нулевую производную по времени, то есть сами функции состояния или просто состояния, 1-ю производную по времени, то есть скорость изменения во времени этих функций, 2-ю производную по времени и т. д. В этом случае как наблюдаемые состояния s, так и ненаблюдаемые состояния υ предстают как наборы из производных по времени разных порядков. Такие наборы производных представляют собой векторы. Под вектором здесь мы понимаем набор величин, над которым производятся некоторые операции. Векторы в подавляющем большинстве наших формул будут, для экономии места, представлены в форме вектор-строк с использованием верхнего индекса T.

Набор из производных разного порядка по времени от функций, описывающих наблюдаемые состояния s, мы будем называть *обобщёнными сенсорными состояниями* и обозначать $\boldsymbol{s}$ или

$$\mathbf{s} = [s, \dot{s}, \ddot{s}]^T, \tag{24}$$

где $\mathbf{s}$ — наблюдаемые состояния сенсорных рецепторов, $\dot{s}$ — 1-я производная по времени от наблюдаемых состояний или скорость изменения наблюдаемых состояний, $\ddot{s}$ — 2-я производная по времени от наблюдаемых состояний или ускорение изменения наблюдаемых состояний [1]-[12].

Несмотря на то, что обобщённые наблюдаемые состояния согласно Фристону [3] представляют собой, в общем случае, бесконечномерный вектор, в практических расчетах такой вектор всегда используется в конечном виде. Количество производных по времени в этом конечномерном векторе Фристон [89] называет *порядком вложения* (emdedding order) или *динамическим порядком*. Например, если порядок вложения равен 2, то такой вектор содержит всего 3 элемента: саму функцию, определяющую состояние, а также первую и вторую производную по времени этой функции. Мы в нашей статье будем использовать порядок вложения равный 2.

Использование слова «обобщённые» связано с тем, что такой набор из производных по времени разных порядков — это и есть обобщённые координаты; соответственно, наблюдаемые состояния, представленные с использованием обобщённых координат — это обобщённые наблюдаемые состояния.

Аналогично наблюдаемым состояниям, ненаблюдаемые состояния, представленные в форме обобщённых координат — это обобщённые ненаблюдаемые состояния. Последние тоже являются вектором. Обозначаться обобщённые ненаблюдамые состояния будут аналогично наблюдаемым — $\boldsymbol{\upsilon}$, и представляют из себя набор из производных разного порядка по времени от ненаблюдаемых состояний, то есть

$$\boldsymbol{\upsilon} = [\upsilon, \dot{\upsilon}, \ddot{\upsilon}]^T, \tag{25}$$

где $\boldsymbol{\upsilon}$ — ненаблюдаемые состояния, $\dot{\upsilon}$— 1-я производная по времени от ненаблюдаемых состояний, или скорость изменения ненаблюдаемых состояний, $\ddot{\upsilon}$— 2-я производная по времени от ненаблюдаемых состояний, или ускорение изменения ненаблюдаемых состояний.

Для того чтобы завершить переход в обобщённые координаты, во всех уравнениях, приведённых выше, s заменяется на $\boldsymbol{s}$, υ на $\boldsymbol{\upsilon}$, все остальные обозначения остаются теми же, а уравнения 17, 23, которые мы будем использовать в качестве основы в дальнейшем, принимают следующий вид:

$$\begin{aligned} F(\boldsymbol{s}, q(\boldsymbol{\upsilon})) &= \int q(\boldsymbol{\upsilon}) \ln \frac{q(\boldsymbol{\upsilon})}{p(\boldsymbol{s}, \boldsymbol{\upsilon})} d\boldsymbol{\upsilon} \\ &= -\int q(\boldsymbol{\upsilon}) \ln p(\boldsymbol{s}, \boldsymbol{\upsilon}) d\boldsymbol{\upsilon} + \int q(\boldsymbol{\upsilon}) \ln q(\boldsymbol{\upsilon}) d\boldsymbol{\upsilon} \\ &= L(\boldsymbol{s}, \boldsymbol{\upsilon}) - H(q(\boldsymbol{\upsilon})). \end{aligned} \tag{26}$$

И следовательно, в дальнейшем наблюдаемые и ненаблюдаемые состояния используются только в форме обобщённых координат, то есть $\boldsymbol{s}$ и $\boldsymbol{\upsilon}$, соответственно.

Во-вторых, нетрудно заметить, что сама ИСЭ в форме уравнения 26, так и величины её слагающие, а именно информационная энергия Гиббса и дифференциальная энтропия, являются интегральными величинами. Помимо этого уравнение 26 содержит неизвестную и принципиально недоступную для мозга, по соображениям изложенным выше, величину $\boldsymbol{\upsilon}$. Это уже само по себе, мягко говоря, существенно осложняет расчет этих величин. Однако уравнение 26 можно существенно упростить. Для этого Фристон использует ***приближение Лапласа*** [2], [3], [5], [7], [8].

Величина $\boldsymbol{\upsilon}$, соответствующая обобщённым ненаблюдаемым состояниям — это то, как уже было сказано, что неизвестно мозгу. Однако

для того чтобы рассчитать свободную энергию и, что эквивалентно, выполнить байесовский вывод, эту величину необходимо каким-либо путём найти. Фристон предполагает [2], [3], [5], [7], [8], что *мозг делает гипотезы о математическом ожидании ненаблюдаемых состояний.* Или другими словами, *мозг кодирует условное математическое ожидание ненаблюдаемых состояний.* Иначе говоря, *математическое ожидание неизвестной величины* $\boldsymbol{\upsilon}$*, которое тоже неизвестно, как и сами ненаблюдаемые состояния, приближается величиной* $\boldsymbol{\mu}_{\upsilon}$. Так как теория Фристона сформулирована в обобщённых координатах, то величина $\boldsymbol{\mu}_{\upsilon}$ также предстает в обобщённых координатах. В этом случае вариационное распределение $q(\boldsymbol{\upsilon})$ заменяется на $q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})$ и становится функцией от состояний среды и гипотез, которые делает мозг о математическом ожидании ненаблюдаемых состояний. Точка с запятой в $q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})$ означает, что речь идёт не об условной зависимости $\boldsymbol{\mu}_{\upsilon}$ от $\boldsymbol{\upsilon}$, а о том, что, говоря языком байесовской теории вероятностей, $\boldsymbol{\upsilon}$ параметризуется $\boldsymbol{\mu}_{\upsilon}$. Аналогичным образом дифференциальная энтропия вариационного распределения $H[q(\boldsymbol{\upsilon})]$ преобразуется в $H[q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})]$. О форме вариационного распределения $q(\boldsymbol{\upsilon})$ и, соответственно, $q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})$, исходно ничего не известно, поэтому *Фристон* [2], [3], [5], [7], [8] *делает допущение, что эта величина имеет форму нормального или гауссова распределения*, или $q(\boldsymbol{\upsilon}) \approx q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon}) = N(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon},C)$. В вариационных байесовских методах такой подход, при котором вариационное распределение приближается нормальным, соответствует *приближению Лапласа* (Приближение Лапласа, также известное как метод перевала используется не только в вариационных байесовских методах, но и при решении разных физических задач [90]. Более подробно о использовании приближения Лапласа в теории Фристона смотри в работах [2], [8].). При этом приближении дифференциальная энтропия вариационного распределения $H[q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})]$ параметризованная $\boldsymbol{\mu}_{\upsilon}$ будет выглядеть следующим образом, причём, как будет видно из дальнейшего, замену вариационного распределения достаточно провести только под знаком логарифма:

$$H[q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})] = -\int q(\boldsymbol{\upsilon}) \ln N(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon},C) d\boldsymbol{\upsilon} \tag{27}$$

или:

$$H[q(\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon})] = -\int q(\boldsymbol{\upsilon})(-\frac{1}{2}(\boldsymbol{\upsilon}-\boldsymbol{\mu}_{\upsilon})C^{-1}(\boldsymbol{\upsilon}-\boldsymbol{\mu}_{\upsilon})^T - \frac{1}{2}\ln|C| - \frac{n}{2}\ln(2\pi))d\boldsymbol{\upsilon}. \tag{28}$$

где n — размерность $\boldsymbol{\mu}_{\upsilon}$. Выражение 28 можно существенно упростить.

Матрица ковариации для данного многомерного нормального распределения, по определению [91], определяется следующим выражением:

$$C = \frac{1}{(2\pi)^{n/2}|C|^{1/2}} \int e^{-\frac{1}{2}(\boldsymbol{\upsilon}-\boldsymbol{\mu}_{\upsilon})C^{-1}(\boldsymbol{\upsilon}-\boldsymbol{\mu}_{\upsilon})^T} (\boldsymbol{\upsilon}-\boldsymbol{\mu}_{\upsilon})(\boldsymbol{\upsilon}-\boldsymbol{\mu}_{\upsilon})^T d\boldsymbol{\upsilon}. \tag{29}$$

Учитывая уравнение 29 и условие нормировки, то есть уравнение 10, дифференциальную энтропию можно представить следующим образом:

$$H[\boldsymbol{\upsilon};\boldsymbol{\mu}_{\upsilon}] = \frac{1}{2} + \frac{1}{2}\ln|C| + \frac{n}{2}\ln(2\pi). \tag{30}$$

Введём, вслед за Фристоном [2], [3], [5], [7], [8], величину, называемую *закодированная по Лапласу информационная энергия* $U(\boldsymbol{s},\boldsymbol{\upsilon})$, равную натуральному логарифму от совместного распределения наблюдаемых и ненаблюдаемых состояний.

$$U(\boldsymbol{s},\boldsymbol{\upsilon}) = \ln p(\boldsymbol{s},\boldsymbol{\upsilon}) \tag{31}$$

Разложим $U(\boldsymbol{s},\boldsymbol{\upsilon})$ в ряд Тейлора по $\boldsymbol{\upsilon}$ в $\boldsymbol{\mu}_{\upsilon}$, проинтегрируем по $d\boldsymbol{\upsilon}$, отбросим производные порядка выше 2 и тогда, учитывая то, что второй член ряда равен нулю, потому что интеграл равен математическому ожиданию, в результате получим:

$$L(\boldsymbol{s},\boldsymbol{\mu}_{\upsilon}) = U(\boldsymbol{s},\boldsymbol{\mu}_{\upsilon}) + \frac{1}{2}C\,\frac{\partial^2}{\partial\boldsymbol{\upsilon}^2}U(\boldsymbol{s},\boldsymbol{\upsilon})\bigg|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_{\upsilon}}. \tag{32}$$

В этом случае ИСЭ принимает вид:

$$F(\boldsymbol{s},\boldsymbol{\mu}_{\upsilon}) = U(\boldsymbol{s},\boldsymbol{\mu}_{\upsilon}) + \frac{1}{2}C\,\frac{\partial^2}{\partial\boldsymbol{\upsilon}^2}U(\boldsymbol{s},\boldsymbol{\upsilon})\bigg|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_{\upsilon}} - \frac{1}{2} - \frac{1}{2}\ln|C| - \frac{n}{2}\ln(2\pi). \tag{33}$$

Фристон [7] пишет о том, что если минимизировать информационное свободное действие и, соответственно, ИСЭ относительно условных обратных ковариаций (условных точностей), то есть посмотреть, в каком случае частная производная ИСЭ по обратной ковариации равна нулю, то выяснится, что обратная ковариация является аналитической функцией от математического ожидания. Другими словами, для выполнения минимизации ИСЭ достаточно того, чтобы она зависела только от математического ожидания ненаблюдаемых состояний и сенсорных входов, то есть

$$\begin{aligned}
&\frac{\partial}{\partial C}F(\boldsymbol{s},\boldsymbol{\mu}_{\upsilon}) = \\
&= \frac{\partial}{\partial C}\left(U(\boldsymbol{s},\boldsymbol{\mu}_{\upsilon}) + \frac{1}{2}C\,\frac{\partial^2}{\partial\boldsymbol{\upsilon}^2}U(\boldsymbol{s},\boldsymbol{\upsilon})\bigg|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_{\upsilon}} - \frac{1}{2} - \frac{1}{2}\ln|C| - \frac{n}{2}\ln(2\pi)\right) \\
&\frac{\partial}{\partial C}\left(\frac{1}{2}C\,\frac{\partial^2}{\partial\boldsymbol{\upsilon}^2}U(\boldsymbol{s},\boldsymbol{\upsilon})\bigg|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_{\upsilon}} - \frac{1}{2}\ln|C|\right) = 0 \\
&\frac{1}{2}\,\frac{\partial^2}{\partial\boldsymbol{\upsilon}^2}U(\boldsymbol{s},\boldsymbol{\upsilon})\bigg|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_{\upsilon}} - \frac{1}{2}C^{-1} = 0 \\
&\frac{\partial^2}{\partial\boldsymbol{\upsilon}^2}U(\boldsymbol{s},\boldsymbol{\upsilon})\bigg|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_{\upsilon}} = C^{-1}.
\end{aligned} \tag{34}$$

Подставив выражение 34 в уравнение 33 и отбросив константу $\frac{n}{2}\ln(2\pi)$, получим значительное упрощение выражения:

$$F(\boldsymbol{s}, \boldsymbol{\mu}_\upsilon) = U(\boldsymbol{s}, \boldsymbol{\mu}_\upsilon) + \frac{1}{2}\ln\left|\frac{\partial^2}{\partial \boldsymbol{\upsilon}^2} U(\boldsymbol{s}, \boldsymbol{\upsilon})\Big|_{\boldsymbol{\upsilon}=\boldsymbol{\mu}_\upsilon}\right|. \tag{35}$$

Нет никакой разницы в том, сначала ли продифференцировать по $\boldsymbol{\upsilon}$ закодированную по Лапласу информационную энергию, а потом заменить в получившемся тождестве $\boldsymbol{\upsilon}$ на $\boldsymbol{\mu}_\upsilon$, или сразу заменить $\boldsymbol{\upsilon}$ на $\boldsymbol{\mu}_\upsilon$, а потом продифференцировать по $\boldsymbol{\mu}_\upsilon$. С учётом этого уравнение 35 можно сразу переписать как:

$$F(\boldsymbol{s}, \boldsymbol{\mu}_\upsilon) = U(\boldsymbol{s}, \boldsymbol{\mu}_\upsilon) + \frac{1}{2}\ln\left|\frac{\partial^2}{\partial \boldsymbol{\mu}_\upsilon^2} U(\boldsymbol{s}, \boldsymbol{\mu}_\upsilon)\right|. \tag{36}$$

Как видно из уравнения 36, при принятии *гипотезы о кодировании мозгом математического ожидания ненаблюдаемых состояний и допущения о гауссовой форме вариационного распределения* можно обнаружить 2 вещи. Во-первых, оказывается, что ИСЭ зависит только от плотности совместного распределения сенсорных данных и гипотез о математическом ожидании ненаблюдаемых состояний (порождающей модели). Другими словами, при принятии этих гипотез вместо принципиально, в рамках фристоновского подхода, неизвестных и недоступных для мозга ненаблюдаемых состояний, для расчёта ИСЭ используются доступные и наблюдаемые величины, что и позволяет её рассчитать. А во-вторых, для того, чтобы вычислить ИСЭ достаточно задать лишь совместное распределение сенсорных данных и гипотез о математическом ожидании ненаблюдаемых состояний.

Вспомним положение из 1-ой главы о том, что ненаблюдаемые состояния слагаются из скрытых (латентных) состояний x, скрытых причин ν и параметров состояний внешней среды θ и γ. Соответственно гипотезы о математическом ожидании ненаблюдаемых состояний $\boldsymbol{\mu}_\upsilon$ будут состоять из гипотез о математическом ожидании скрытых состояний $\boldsymbol{\mu}_x$, скрытых причин $\boldsymbol{\mu}_\nu$, параметров $\boldsymbol{\mu}_\theta$ и гиперпараметров $\boldsymbol{\mu}_\gamma$. Термин «гиперпараметры» как у Фристона [2], [3], [5], [7], [8], так и в байесовских методах машинного обучения, используется для того, чтобы как-то различать параметры θ и γ и, соответственно, $\boldsymbol{\mu}_\theta$ и $\boldsymbol{\mu}_\gamma$.

Другими словами, $\boldsymbol{\mu}_\upsilon$ можно рассматривать как вектор, состоящий из $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_\nu$, $\boldsymbol{\mu}_\theta$ и $\boldsymbol{\mu}_\gamma$ или

$$\boldsymbol{\mu}_\upsilon = [\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma]^T. \tag{37}$$

Тогда уравнение 36 преобразуется к следующему виду:

$$F(\boldsymbol{s}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) = U(\boldsymbol{s}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) + \frac{1}{2}\ln\left|\frac{\partial^2}{\partial \boldsymbol{\mu}_\upsilon^2} U(\boldsymbol{s}, \boldsymbol{\mu}_\upsilon)\right|. \tag{38}$$

В этом случае вторую вариационную производную от закодированной по Лапласу информационной энергии в уравнении 38 Фристон [7] представляет аналогично случаю, когда векторный оператор Лапласа действует на вектор, то есть в форме квадратной матрицы следующего вида:

$$\frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\upsilon^2} = \begin{pmatrix} \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_x^2} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_x \partial \boldsymbol{\mu}_\nu} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_x \partial \boldsymbol{\mu}_\theta} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_x \partial \boldsymbol{\mu}_\gamma} \\ \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\nu \partial \boldsymbol{\mu}_x} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\nu^2} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\nu \partial \boldsymbol{\mu}_\theta} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\nu \partial \boldsymbol{\mu}_\gamma} \\ \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\theta \partial \boldsymbol{\mu}_x} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\theta \partial \boldsymbol{\mu}_\nu} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\theta^2} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\theta \partial \boldsymbol{\mu}_\gamma} \\ \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\gamma \partial \boldsymbol{\mu}_x} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\gamma \partial \boldsymbol{\mu}_\nu} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\gamma \partial \boldsymbol{\mu}_\theta} & \frac{\partial^2 U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon)}{\partial \boldsymbol{\mu}_\gamma^2} \end{pmatrix}. \tag{39}$$

Закодированная по Лапласу информационная энергия $U(\boldsymbol{s},\boldsymbol{\upsilon})$ тогда преобразуется в $U(\boldsymbol{s},\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$ и задаётся следующим уравнением:

$$\begin{aligned} U(\boldsymbol{s},\boldsymbol{\mu}_\upsilon) &= U(\boldsymbol{s},\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma) \\ &= \ln p(\boldsymbol{s}|\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma) + \ln p(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma) \\ &= \ln p(\boldsymbol{s}|\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma) + \ln p(\boldsymbol{\mu}_x|\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma) \\ &+ \ln p(\boldsymbol{\mu}_\nu) + \ln p(\boldsymbol{\mu}_\theta) + \ln p(\boldsymbol{\mu}_\gamma). \end{aligned} \tag{40}$$

Аналогично случаю с вариационным распределением, как правдоподобие $p(\boldsymbol{s}|\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$, так и условное распределение $p(\boldsymbol{\mu}_x|\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$, а также априорные вероятности $p(\boldsymbol{\mu}_\nu), p(\boldsymbol{\mu}_\theta), p(\boldsymbol{\mu}_\gamma)$, приближаются Фристоном [7] гауссовыми распределениями, что, опять же, составляет приближение Лапласа. Единственное различие здесь состоит в том, что для правдоподобия $p(\boldsymbol{s}|\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$ вместо $\boldsymbol{\mu}_\upsilon$ используется $g(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$. А для условного распределения $p(\boldsymbol{\mu}_x|\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$ вместо математического ожидания случайной величины используется $f(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta,\boldsymbol{\mu}_\gamma)$. А для самой случайной величины используется математическое ожидание, умноженное на квадратную матрицу D, т.е. $D\boldsymbol{\mu}_x$. В общем случае матрица D имеет следующий вид:

$$D = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{41}$$

Так как мы используем порядок вложения (динамический порядок) равный 2, то математическое ожидание скрытых состояний в обобщённых координатах $\boldsymbol{\mu}_x$ принимает такую форму:

$$\boldsymbol{\mu}_x = [\mu_x, \dot{\mu}_x, \ddot{\mu}_x]^T \tag{42}$$

Тогда $D\boldsymbol{\mu}_x$ выглядит следующим образом:

$$D\boldsymbol{\mu}_x = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_x \\ \dot{\mu}_x \\ \ddot{\mu}_x \end{pmatrix} = \begin{pmatrix} \dot{\mu}_x \\ \ddot{\mu}_x \\ 0 \end{pmatrix} \tag{43}$$

В этом случае правдоподобие принимает такую форму:

$$p(\boldsymbol{s}|\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) = \frac{1}{\sqrt{(2\pi)^d|\tilde{C}_z|}} e^{-\frac{1}{2}(s-g(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta))^T \tilde{C}_z^{-1}(s-g(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta))}. \tag{44}$$

А условная вероятность тогда становится:

$$\begin{aligned} p(\boldsymbol{\mu}_x|\boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) &= p(D\boldsymbol{\mu}_x|\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) \\ &= \frac{1}{\sqrt{(2\pi)^d|\tilde{C}_w|}} e^{-\frac{1}{2}(D\boldsymbol{\mu}_x - f(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta))^T \tilde{C}_w^{-1}(D\boldsymbol{\mu}_x - f(\boldsymbol{\mu}_x,\boldsymbol{\mu}_\nu,\boldsymbol{\mu}_\theta))}. \end{aligned} \tag{45}$$

Где матрицы ковариации $\tilde{C}_z$ и $\tilde{C}_w$ в обобщённых координатах в самом общем виде содержат функции, зависящие от гиперпараметров, а также от математических ожиданий скрытых и каузальных состояний [7]. Однако подробностей, касающихся матриц такого вида Фристон в работе [7] не приводит, а наиболее часто использует, в том числе в той же работе более простой вид этих матриц, когда последние зависят только от гиперпараметров γ. Матрица ковариации $\tilde{C}_z$ представляет собой тензорное произведение Кронекера матрицы ковариации $R(\gamma)$ и матрицы ковариации C_z, т.е. $\tilde{C}_z = R(\gamma) \otimes C_z$. Аналогично для матрицы ковариации в обобщённых координатах $\tilde{C}_w$, т.е. $\tilde{C}_w = R(\gamma) \otimes C_w$

$$C_z = \begin{pmatrix} e^{-\lambda_z} & 0 & 0 \\ 0 & e^{-\lambda_z} & 0 \\ 0 & 0 & e^{-\lambda_z} \end{pmatrix}, \tag{46}$$

$$C_w = \begin{pmatrix} e^{-\lambda_w} & 0 & 0 \\ 0 & e^{-\lambda_w} & 0 \\ 0 & 0 & e^{-\lambda_w} \end{pmatrix}, \tag{47}$$

$$R(\gamma) = \begin{pmatrix} 1 & 0 & -\frac{1}{2}\gamma \\ 0 & \frac{1}{2}\gamma & 0 \\ -\frac{1}{2}\gamma & 0 & \frac{3}{4}\gamma \end{pmatrix}, \tag{48}$$

Априорное распределение $p(\boldsymbol{\mu}_\nu)$ это:

$$p(\boldsymbol{\mu}_\nu) = \frac{1}{\sqrt{(2\pi)^d|\tilde{C}_\nu|}} e^{-\frac{1}{2}(\boldsymbol{\mu}_\nu - \boldsymbol{\eta}_\nu)^T \tilde{C}_\nu^{-1}(\boldsymbol{\mu}_\nu - \boldsymbol{\eta}_\nu)}. \tag{49}$$

Априорное распределение параметров $p(\boldsymbol{\mu}_\theta)$ это:

$$p(\boldsymbol{\mu}_\theta) = \frac{1}{\sqrt{(2\pi)^d|\tilde{C}_\theta|}} e^{-\frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\eta}_\theta)^T \tilde{C}_\theta^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\eta}_\theta)}. \tag{50}$$

Априорное распределение гиперпараметров $p(\boldsymbol{\mu}_\gamma)$ это:

$$p(\boldsymbol{\mu}_\gamma) = \frac{1}{\sqrt{(2\pi)^d |\tilde{C}_\gamma|}} e^{-\frac{1}{2}(\boldsymbol{\mu}_\gamma - \boldsymbol{\eta}_\gamma)^T \tilde{C}_\gamma^{-1} (\boldsymbol{\mu}_\gamma - \boldsymbol{\eta}_\gamma)}. \tag{51}$$

Элементы матриц ковариации C_θ и C_γ исходно задаются как нулевые, в процессе реализации алгоритма динамической максимизации математического ожидания (DEM) в этих матрицах происходит замена нулевых элементов на ненулевые, что легко заметить запустив любую модель из пакета SPM [89].

Тогда закодированная по Лапласу энергия, отбрасывая константы $-\frac{d}{2}\ln(2\pi)$ записывается как:

$$\begin{aligned}
&U(\boldsymbol{s}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) = \\
&= -\frac{1}{2}(s - g(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta))^T \tilde{C}_z^{-1} (s - g(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta)) - \frac{1}{2}\ln|\tilde{C}_z| \\
&- \frac{1}{2}(D\boldsymbol{\mu}_x - f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta))^T \tilde{C}_w^{-1} (D\boldsymbol{\mu}_x - f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta)) - \frac{1}{2}\ln|\tilde{C}_w| \\
&- \frac{1}{2}(\boldsymbol{\mu}_\nu - \boldsymbol{\eta}_\nu)^T \tilde{C}_\nu^{-1} (\boldsymbol{\mu}_\nu - \boldsymbol{\eta}_\nu) - \frac{1}{2}\ln|\tilde{C}_\nu| - \frac{1}{2}(\boldsymbol{\mu}_\theta - \boldsymbol{\eta}_\theta)^T \tilde{C}_\theta^{-1} (\boldsymbol{\mu}_\theta - \boldsymbol{\eta}_\theta) \\
&- \frac{1}{2}\ln|\tilde{C}_\theta| - \frac{1}{2}(\boldsymbol{\mu}_\gamma - \boldsymbol{\eta}_\gamma)^T \tilde{C}_\gamma^{-1} (\boldsymbol{\mu}_\gamma - \boldsymbol{\eta}_\gamma) - \frac{1}{2}\ln|\tilde{C}_\gamma|.
\end{aligned} \tag{52}$$

Для того чтобы вычислить вторую вариационную производную от закодированной по Лапласу информационной энергии, в уравнении нужно продифференцировать правую часть уравнения 52 так, как это показано в уравнении 39. В результате опять-таки получится матрица. За более подробной информации об элементах этой матрицы, отсылаем читателя к работе [7].

Для s-$g(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma)$ и $D\boldsymbol{\mu}_x$-$f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma)$ Фристон [2],[3],[5],[7],[8] вводит следующие обозначения:

$$\boldsymbol{\varepsilon}_\nu = \boldsymbol{s}\text{-}g(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma), \tag{53}$$

$$\boldsymbol{\varepsilon}_x = D\boldsymbol{\mu}_x\text{-}f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma). \tag{54}$$

где $\boldsymbol{\varepsilon}_\nu$ — обобщённая ошибка предсказания сенсорных состояний, $\boldsymbol{\varepsilon}_x$ — обобщённая ошибка предсказания движения скрытых состояний. Опять же обе эти величины представлены в форме обобщённых координат. Величина $g(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma)$ — это *предсказание наблюдаемых* (в данном случае сенсорных) *состояний*, а величина $f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma)$ — это *предсказание ненаблюдаемых состояний* (в данном случае скрытых состояний среды).

В результате уравнение 52 может быть представлено в более компактной форме:

$$\begin{aligned} U(\boldsymbol{s}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma) = &- \frac{1}{2}\boldsymbol{\varepsilon}_\nu^T \tilde{C}_z^{-1}\boldsymbol{\varepsilon}_\nu - \frac{1}{2}\ln|\tilde{C}_z| - \frac{1}{2}\boldsymbol{\varepsilon}_x^T \tilde{C}_w^{-1}\boldsymbol{\varepsilon}_x - \frac{1}{2}\ln|\tilde{C}_w| \\ &- \frac{1}{2}(\boldsymbol{\mu}_\nu - \boldsymbol{\eta}_\nu)^T \tilde{C}_\nu^{-1}(\boldsymbol{\mu}_\nu - \boldsymbol{\eta}_\nu) - \frac{1}{2}\ln|\tilde{C}_\nu| \\ &- \frac{1}{2}\boldsymbol{\varepsilon}_\theta^T \tilde{C}_\theta^{-1}\boldsymbol{\varepsilon}_\theta - \frac{1}{2}\ln|\tilde{C}_\theta| - \frac{1}{2}\boldsymbol{\varepsilon}_\gamma^T \tilde{C}_\gamma^{-1}\boldsymbol{\varepsilon}_\gamma - \frac{1}{2}\ln|\tilde{C}_\gamma|. \end{aligned} \tag{55}$$

Здесь нужно напомнить, что следует различать обобщённые скрытые и каузальные состояния мира: x и ν соответственно, их параметры θ, и гиперпараметры γ и гипотезы о математическом ожидании скрытых и каузальных состояниях мира $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_\nu$, и параметров $\boldsymbol{\mu}_\theta$, и гиперпараметров $\boldsymbol{\mu}_\gamma$ этих гипотез.

Скрытые состояния мира, согласно Фристону [1]-[12], имеют динамику, описываемую дифференциальным уравнением следующего вида:

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{a}) + \boldsymbol{w}, \tag{56}$$

где скорость изменения скрытых состояний является суммой функции от скрытых и каузальных состояний, их параметров и гиперпараметров, действия (у Фристона употребляется термин «action») и шума. А сенсорные состояния представляют собой сумму функции от скрытых состояний и причин, их параметров и гиперпараметров, действия (action) и шума. Под «действием» (не путать с информационным свободным действием) здесь понимается некоторая переменная связанная с моторной активностью. Мы предлагаем понимать «действие» («action») в более широком смысле — как деятельность организма в среде, приводящую к изменениям в ней, и включающую в себя не только двигательную активность, но и секреторную, и экскреторную.

Другими словами, скрытые, каузальные состояния и их параметры, и гиперпараметры, и действие — это аргумент, а сенсорное состояние — это результат действия некоторой функции на аргумент плюс шум, то есть

$$\boldsymbol{s} = g(\boldsymbol{x}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{a}) + \boldsymbol{z}. \tag{57}$$

Теперь, если сравнить уравнения 56, 57 и, например, уравнение 53, найдя в последнем уравнении члены: $\boldsymbol{s}$-$g(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma)$ и $D\boldsymbol{\mu}_x$-$f(\boldsymbol{\mu}_x, \boldsymbol{\mu}_\nu, \boldsymbol{\mu}_\theta, \boldsymbol{\mu}_\gamma)$ можно легко заметить очевидное сходство. Однако это хоть и схожие, но разные величины, и их следует различать. Тем не менее, это сходство говорит о большой роли шума, которую он играет в теории Фристона. Об этом же говорит и сходство уравнений 56 и 57 с уравнением Ланжевена.

Кроме того, заметим, что аналитическое решение уравнения Ланжевена для броуновской частицы совпадает с гауссианом. Действительно работы Фристона подтверждают эту роль [3], [5]. Так как вся теория Фристона сформулирована с использованием обобщённых координат, то если распаковать уравнения 56 и 57, в *предположении локальной линейности* [1]-[12] они будут выглядеть так:

$$
\begin{aligned}
s &= g(x,\nu,\theta,\gamma,a) + z \\
\dot{s} &\approx \dot{x}\frac{\partial}{\partial x}g(x,\nu,\theta,\gamma,a) + \dot{\nu}\frac{\partial}{\partial \nu}g(x,\nu,\theta,\gamma,a) + \dot{z} \\
\ddot{s} &\approx \ddot{x}\frac{\partial}{\partial x}g(x,\nu,\theta,\gamma,a) + \ddot{\nu}\frac{\partial}{\partial \nu}g(x,\nu,\theta,\gamma,a) + \ddot{z} \\
\dot{x} &= f(x,\nu,\theta,\gamma,a) + w \\
\ddot{x} &\approx \dot{x}\frac{\partial}{\partial x}f(x,\nu,\theta,\gamma,a) + \dot{\nu}\frac{\partial}{\partial \nu}f(x,\nu,\theta,\gamma,a) + \dot{w} \\
\dddot{x} &\approx \ddot{x}\frac{\partial}{\partial x}f(x,\nu,\theta,\gamma,a) + \ddot{\nu}\frac{\partial}{\partial \nu}f(x,\nu,\theta,\gamma,a) + \ddot{w}.
\end{aligned}
\tag{58}
$$

Предположение локальной линейности состоит в том, что нелинейные члены в уравнении 58 отбрасываются, что опять-таки сильно упрощает расчёты. Аналогичным образом предстаёт и уравнение 59.

$$
\begin{aligned}
\varepsilon_\nu &= \dot{\mu}_\nu - g(\mu_x,\mu_\nu) \\
\dot{\varepsilon}_\nu &\approx \ddot{\mu}_\nu - \dot{\mu}_x\frac{\partial}{\partial \mu_x}g(\mu_x,\mu_\nu) - \dot{\mu}_\nu\frac{\partial}{\partial \mu_\nu}g(\mu_x,\mu_\nu) \\
\ddot{\varepsilon}_\nu &\approx \dddot{\mu}_\nu - \ddot{\mu}_x\frac{\partial}{\partial \mu_x}g(\mu_x,\mu_\nu) - \ddot{\mu}_\nu\frac{\partial}{\partial \mu_\nu}g(\mu_x,\mu_\nu) \\
\varepsilon_x &= \dot{\mu}_x - f(\mu_x,\mu_\nu) \\
\dot{\varepsilon}_x &\approx \ddot{\mu}_x - \dot{\mu}_x\frac{\partial}{\partial \mu_x}f(\mu_x,\mu_\nu) - \dot{\mu}_\nu\frac{\partial}{\partial \mu_\nu}f(\mu_x,\mu_\nu) \\
\ddot{\varepsilon}_x &\approx \dddot{\mu}_x - \ddot{\mu}_x\frac{\partial}{\partial \mu_x}f(\mu_x,\mu_\nu) - \ddot{\mu}_\nu\frac{\partial}{\partial \mu_\nu}f(\mu_x,\mu_\nu).
\end{aligned}
\tag{59}
$$

Подводя итог в отношении использования Фристоном приближения Лапласа, как дополнительной гипотезы, можно сказать, что её использование позволяет свести расчёт ИСЭ к достаточно громоздким (смотри вторую производную закодированной по Лапласу энергии) уравнениям, однако, в отличие от исходного интегрального представления ИСЭ, они поддаются расчёту.

Итак, в этом разделе нами была рассмотрена простейшая ситуация, описывающая взаимодействие среды, сенсорных рецепторов и мозга. Была введена гипотеза байесовского мозга во фристоновской формулировке

и установлена связь между ИСЭ, введенной в предыдущем разделе, и гипотезой байесовского мозга. Однако хорошо известно, что мозг состоит из разных отделов, различающихся как анатомически, так и функционально, физиологически. В следующем разделе эта простейшая ситуация взаимодействия среды, рецепторов и мозга будет нами существенно расширена и дополнена за счёт введения третьей важной концепции теории Фристона — «предиктивного кодирования» (во фристоновской формулировке), показывающей взаимоотношения между разными отделами мозга.

4. Фристоновская формулировка гипотезы предиктивного кодирования

В этом разделе мы рассматриваем систему, состоящую из среды, сенсорных рецепторов и разных отделов мозга, в первую очередь коры больших полушарий. По аналогии со схемой, изложенной в предыдущем разделе, где состояния среды в обобщённых координатах являются обобщёнными ненаблюдаемыми состояниями, а сенсорные состояния в обобщённых координатах являются обобщёнными наблюдаемыми состояниями, где в роли «наблюдателя» выступает мозг, также и отделы мозга могут быть представлены в подобных взаимоотношениях. Для начала примем, что в этом случае состояния каждого отдела мозга описываются в обобщённых координатах двумя группами состояний: одни из этих состояний будут обобщёнными наблюдаемыми, так как они могут «наблюдаться» вышележащими отделами мозга, а другие состояния — обобщёнными ненаблюдаемыми, так как они не могут непосредственно наблюдаться вышележащими отделами мозга. Разумеется, обобщённые ненаблюдаемые состояния нижележащих структур влияют на обобщённые наблюдаемые состояния, а также на другие обобщённые ненаблюдаемые состояния внутри своего отдела мозга. Соответственно, «наблюдателями» являются вышележащие отделы мозга, напрямую зависящие от наблюдаемых состояний нижележащих отделов мозга.

Согласно Фристону [1], [123], такая схема удачно накладывается на архитектуру мозга. В известной работе Felleman и Van Essen (1991) [92], авторы показали, что зрительная и моторная системы макаки организованы иерархически. Так иерархия зрительной кортикальной системы, согласно этим авторам, содержит 10 уровней. Всего в иерархии зрительной системы, исходя из содержания статьи этих авторов, можно насчитать 14 уровней, если включить в эту систему ещё сетчатку, латеральное коленчатое ядро, как самые нижние уровни иерархии, а также энторинальную кору и гиппокамп — как самые высокие уровни иерархии. Стоит заме-

тить, что авторы работы рассматривали главные, «магистральные» пути зрительной системы макаки, поэтому в их схему не попали, например, верхнее двухолмие, подушка таламуса и супрахиазмальное ядро гипоталамуса.

Применяя этот подход с ненаблюдаемыми и наблюдаемыми состояниями к иерархии зрительной системы, рассмотренной в работе Felleman и Van Essen, получаем следующую картину. На самом нижнем уровне иерархии зрительной системы находится сетчатка. Состояния сетчатки в обобщённых координатах, аналогично принципу, описанному в предыдущем разделе, мы называем обобщёнными наблюдаемыми состояниями. Эти состояния, по-прежнему, зависят от состояний среды или обобщённых ненаблюдаемых состояний. А вот в роли непосредственного наблюдателя выступает уже не мозг в целом, а только один из его отделов, с которым непосредственно контактирует сетчатка.

Поднимаясь с самого нижнего уровня иерархии зрительной системы, представленной Felleman и Van Essen, то есть от сетчатки, на одну ступень вверх по иерархии, мы попадём в латеральное коленчатое тело. С латеральным коленчатым телом, выступающим по отношению к обобщённым ненаблюдаемым состояниям среды и обобщённым наблюдаемым состояниям сетчатки в роли наблюдателя, могут сопрягаться его собственные обобщённые наблюдаемые состояния, где в роли «наблюдателя» его состояний выступает первичная зрительная кора (V1). А в роли обобщённых ненаблюдаемых состояний, по отношению к первичной зрительной коре, выступают состояния сетчатки, а также состояния ядер таламуса, не связанных прямой связью с этой областью зрительной коры.

Аналогичным образом может быть рассмотрена не только первичная зрительная кора (V1), но и другие уровни иерархии зрительной системы или другие отделы мозга [1]-[12].

К каким же выводам мы приходим, рассматривая иерархическую организацию обработки сенсорной информации в мозге? Обобщая вышесказанное, мы можем сказать, что, во-первых, с каждым из уровней иерархии зрительной системы могут связываться обобщённые наблюдаемые состояния для вышележащего уровня. В-вторых, каждый из этих уровней может выступать в роли «наблюдателя» состояний уровня, расположенного ниже в иерархии. В-третьих, внутри каждого уровня иерархии имеются обобщённые ненаблюдаемые (скрытые от вышележащих уровней) состояния. В-четвертых, каждое обобщённое наблюдаемое состояние i-го уровня иерархии выступает по отношению i+2-му уровню иерархии, то есть следующему через один, как обобщённое ненаблюдаемое состояние [3].

Обобщённые наблюдаемые состояния зрительной системы, представленные в разных её отделах, мы будем называть обобщёнными каузальными состояниями ν. В свою очередь, обобщённые каузальные состояния вместе с обобщёнными наблюдаемыми состояниями сенсорных рецепторов мы будем называть иерархическими обобщёнными наблюдаемыми состояниями.

Обобщённые ненаблюдаемые состояния, включающие в себя обобщённые ненаблюдаемые состояния как среды, так и обобщённые каузальные и обобщённые скрытые состояния разных отделов мозга, мы будем называть иерархическими обобщёнными ненаблюдаемыми состояниями. Обобщённые каузальные состояния мы сюда включили потому, что, как было отмечено выше, они могут выступать также и в роли ненаблюдаемых состояний. Такая двузначность обобщённых каузальных состояний, на первый взгляд, представляет из себя проблему, однако, как мы увидим ниже, сравнивая векторные представления, связанные с обобщёнными каузальными состояниями, никакой проблемы здесь нет, и, что важно, общий вид всех формул сохраняется в прежнем виде. А такое деление на иерархические обобщённые наблюдаемые и ненаблюдаемые состояния — это вопрос договорённости и удобства использования формул.

Иерархические обобщённые наблюдаемые и ненаблюдаемые состояния каждого отдельного уровня иерархии, взятые вместе, мы будем называть модулями состояний или модулями обобщённых состояний [3]. Соответственно, каждому уровню иерархии будет соответствовать свой модуль состояний.

Иерархические обобщённые наблюдаемые состояния мы будем обозначать так же, как и в предыдущей главе, но с использованием нижнего индекса, то есть $\boldsymbol{s}_{(i)}$. Такая запись означает, что мы имеем дело с вектором, состоящим из векторов. Другими словами, иерархические обобщённые наблюдаемые состояния $\boldsymbol{s}_{(i)}$— это вектор, представляющий из себя набор из обобщённых наблюдаемых состояний сенсорных рецепторов и обобщённых каузальных состояний каждого уровня иерархии, которые сами по себе тоже являются векторами или

$$\boldsymbol{s}_{(i)} = [\boldsymbol{s}, \boldsymbol{\mu}_{\nu(1)}, \boldsymbol{\mu}_{\nu(2)}, \boldsymbol{\mu}_{\nu(3)}, ..., \boldsymbol{\mu}_{\nu(n)}]^T, \tag{60}$$

где $\boldsymbol{s}$ — обобщённые наблюдаемые состояния сенсорных рецепторов, $\boldsymbol{\mu}_{\nu(1)}$ — обобщённое каузальное состояние 1-го уровня иерархии, $\boldsymbol{\mu}_{\nu(2)}$ — обобщённое каузальное состояние 2-го уровня иерархии, $\boldsymbol{\mu}_{\nu(3)}$ — обобщённое каузальное состояние 3-го уровня иерархии, $\boldsymbol{\mu}_{\nu(n)}$ — обобщённое каузальное состояние n-го уровня иерархии.

Аналогично иерархическим обобщённым наблюдаемым состояниям, иерархические обобщённые ненаблюдаемые состояния мы будем обозначать так же, как и в предыдущей главе, но будем использовать жирный

шрифт, то есть $\boldsymbol{\mu}_{\nu(i)}$. Иерархические обобщённые ненаблюдаемые состояния $\boldsymbol{\mu}_{\nu(i)}$ — это тоже вектор, представляющий из себя набор из обобщённых скрытых и каузальных состояний среды и обобщённых скрытых и каузальных состояний каждого уровня иерархии, которые сами по себе тоже являются векторами, или

$$\boldsymbol{\mu}_{x(i)} = [\boldsymbol{\mu}_{x(1)}, \boldsymbol{\mu}_{x(2)}, \boldsymbol{\mu}_{x(3)}, ..., \boldsymbol{\mu}_{x(n)}]^T, \tag{61}$$

где $\boldsymbol{\mu}_{x(1)}$ — обобщённое скрытое состояние 1-го уровня иерархии, $\boldsymbol{\mu}_{x(2)}$ — обобщённое скрытое состояние 2-го уровня иерархии, $\boldsymbol{\mu}_{x(3)}$ — обобщённое скрытое состояние 3-го уровня иерархии, $\boldsymbol{\mu}_{x(n)}$ — обобщённое скрытое состояние n-го уровня иерархии;

$$\boldsymbol{\mu}_{\nu(i)} = [\boldsymbol{\mu}_{\nu(1)}, \boldsymbol{\mu}_{\nu(2)}, \boldsymbol{\mu}_{\nu(3)}, ..., \boldsymbol{\mu}_{\nu(n)}]^T, \tag{62}$$

где $\boldsymbol{\mu}_{\nu(1)}$ — обобщённое скрытое состояние 1-го уровня иерархии, $\boldsymbol{\mu}_{\nu(2)}$ — обобщённое скрытое состояние 2-го уровня иерархии, $\boldsymbol{\mu}_{\nu(3)}$ — обобщённое скрытое состояние 3-го уровня иерархии, $\boldsymbol{\mu}_{\nu(n)}$ — обобщённое скрытое состояние n-го уровня иерархии;

$$\boldsymbol{\mu}_{\theta(i)} = [\boldsymbol{\mu}_{\theta(1)}, \boldsymbol{\mu}_{\theta(2)}, \boldsymbol{\mu}_{\theta(3)}, ..., \boldsymbol{\mu}_{\theta(n)}]^T, \tag{63}$$

где $\boldsymbol{\mu}_{\theta(1)}$ — обобщённое скрытое состояние 1-го уровня иерархии, $\boldsymbol{\mu}_{\theta(2)}$ — обобщённое скрытое состояние 2-го уровня иерархии, $\boldsymbol{\mu}_{\theta(3)}$ — обобщённое скрытое состояние 3-го уровня иерархии, $\boldsymbol{\mu}_{\theta(n)}$ — обобщённое скрытое состояние n-го уровня иерархии;

$$\boldsymbol{\mu}_{\lambda(i)} = [\boldsymbol{\mu}_{\lambda(1)}, \boldsymbol{\mu}_{\lambda(2)}, \boldsymbol{\mu}_{\lambda(3)}, ..., \boldsymbol{\mu}_{\lambda(n)}]^T, \tag{64}$$

где $\boldsymbol{\mu}_{\lambda(1)}$ — обобщённое скрытое состояние 1-го уровня иерархии, $\boldsymbol{\mu}_{\lambda(2)}$ — обобщённое скрытое состояние 2-го уровня иерархии, $\boldsymbol{\mu}_{\lambda(3)}$ — обобщённое скрытое состояние 3-го уровня иерархии, $\boldsymbol{\mu}_{\lambda(n)}$ — обобщённое скрытое состояние n-го уровня иерархии.

Тогда, во всех вышеприведенных уравнениях, где фигурирует $\boldsymbol{s}$, эта величина заменяется на $\boldsymbol{s}_{(i)}$, аналогичным образом заменяются и другие величины $\boldsymbol{\mu}_x$ на $\boldsymbol{\mu}_{x(1)}$, $\boldsymbol{\mu}_\nu$ на $\boldsymbol{\mu}_{\nu(1)}$, $\boldsymbol{\mu}_\theta$ на $\boldsymbol{\mu}_{\theta(1)}$, $\boldsymbol{\mu}_\lambda$ на $\boldsymbol{\mu}_{\lambda(1)}$. Каждая матрица ковариации из уравнений 46 и 47 становится первым диагональным элементом главных диагоналей соответствующих блочно-диагональных матриц. Остальные диагональные элементы по виду аналогичны исходным матрицам ковариации из уравнений 46 и 47. Размерность этих блочно-диагональных матриц зависит от количества уровней иерархии порождающей модели. Все остальные обозначения остаются теми же.

При сравнении векторных представлений иерархических обобщённых наблюдаемых и ненаблюдаемых состояний видно, что с каждым n-ным

уровнем иерархии зрительной системы связано n-ное обобщённое скрытое состояние, n-ное обобщённое каузальное состояние и n-1-ое обобщённое каузальное состояние (или для самого нижнего уровня иерархии — сенсорное состояние), а также некоторые параметры и гиперпараметры, связанные с обобщёнными скрытыми и каузальными состояниями [3]. Другими словами, общий вид всех формул сохраняется, подтверждая то, о чём мы писали выше. Состояния иерархии мозга представляются следующим образом, то есть

$$
\begin{aligned}
\varepsilon_{\nu(1)} &= \dot{\mu}_{\nu(1)} - g(\mu_{x(1)}, \mu_{\nu(1)}) \\
\dot{\varepsilon}_{\nu(1)} &\approx \ddot{\mu}_{\nu(1)} - \dot{\mu}_{x(1)} \frac{\partial}{\partial \mu_{x(1)}} g(\mu_{x(1)}, \mu_{\nu(1)}) - \dot{\mu}_{\nu(1)} \frac{\partial}{\partial \mu_{\nu(1)}} g(\mu_{x(1)}, \mu_{\nu(1)}) \\
\ddot{\varepsilon}_{\nu(1)} &\approx \dddot{\mu}_{\nu(1)} - \ddot{\mu}_{x(1)} \frac{\partial}{\partial \mu_{x(1)}} g(\mu_{x(1)}, \mu_{\nu(1)}) - \ddot{\mu}_{\nu(1)} \frac{\partial}{\partial \mu_{\nu(1)}} g(\mu_{x(1)}, \mu_{\nu(1)}) \\
\varepsilon_{x(1)} &= \dot{\mu}_{x(1)} - f(\mu_{x(1)}, \mu_{\nu(1)}) \\
\dot{\varepsilon}_{x(1)} &\approx \ddot{\mu}_{x(1)} - \dot{\mu}_{x(1)} \frac{\partial}{\partial \mu_{x(1)}} f(\mu_{x(1)}, \mu_{\nu(1)}) - \dot{\mu}_{\nu(1)} \frac{\partial}{\partial \mu_{\nu(1)}} f(\mu_{x(1)}, \mu_{\nu(1)}) \\
\ddot{\varepsilon}_{x(1)} &\approx \dddot{\mu}_{x(1)} - \ddot{\mu}_{x(1)} \frac{\partial}{\partial \mu_{x(1)}} f(\mu_{x(1)}, \mu_{\nu(1)}) - \ddot{\mu}_{\nu(1)} \frac{\partial}{\partial \mu_{\nu(1)}} f(\mu_{x(1)}, \mu_{\nu(1)}) \\
\varepsilon_{\nu(2)} &= \dot{\mu}_{\nu(1)} - g(\mu_{x(2)}, \mu_{\nu(2)}) \\
\dot{\varepsilon}_{\nu(2)} &\approx \ddot{\mu}_{\nu(1)} - \dot{\mu}_{x(2)} \frac{\partial}{\partial \mu_{x(2)}} g(\mu_{x(2)}, \mu_{\nu(2)}) - \dot{\mu}_{\nu(2)} \frac{\partial}{\partial \mu_{\nu(2)}} g(\mu_{x(2)}, \mu_{\nu(2)}) \\
\ddot{\varepsilon}_{\nu(2)} &\approx \dddot{\mu}_{\nu(1)} - \ddot{\mu}_{x(2)} \frac{\partial}{\partial \mu_{x(2)}} g(\mu_{x(2)}, \mu_{\nu(2)}) - \ddot{\mu}_{\nu(2)} \frac{\partial}{\partial \mu_{\nu(2)}} g(\mu_{x(2)}, \mu_{\nu(2)}) \\
\varepsilon_{x(2)} &= \dot{\mu}_{x(2)} - f(\mu_{x(2)}, \mu_{\nu(2)}) \\
\dot{\varepsilon}_{x(2)} &\approx \ddot{\mu}_{x(2)} - \dot{\mu}_{x(2)} \frac{\partial}{\partial \mu_{x(2)}} f(\mu_{x(2)}, \mu_{\nu(2)}) - \dot{\mu}_{\nu(2)} \frac{\partial}{\partial \mu_{\nu(2)}} f(\mu_{x(2)}, \mu_{\nu(2)}) \\
\ddot{\varepsilon}_{x(2)} &\approx \dddot{\mu}_{x(2)} - \ddot{\mu}_{x(2)} \frac{\partial}{\partial \mu_{x(2)}} f(\mu_{x(2)}, \mu_{\nu(2)}) - \ddot{\mu}_{\nu(2)} \frac{\partial}{\partial \mu_{\nu(2)}} f(\mu_{x(2)}, \mu_{\nu(2)}).
\end{aligned}
\tag{65}
$$

Уравнения i-го уровня иерархии в предположении локальной линейности

$$
\begin{aligned}
\varepsilon_{\nu(i)} &= \dot{\mu}_{\nu(i-1)} - g(\mu_{x(i)}, \mu_{\nu(i)}) \\
\dot{\varepsilon}_{\nu(i)} &\approx \ddot{\mu}_{\nu(i-1)} - \dot{\mu}_{x(i)} \frac{\partial}{\partial \mu_{x(i)}} g(\mu_{x(i)}, \mu_{\nu(i)}) - \dot{\mu}_{\nu(i)} \frac{\partial}{\partial \mu_{\nu(i)}} g(\mu_{x(i)}, \mu_{\nu(i)}) \\
\ddot{\varepsilon}_{\nu(i)} &\approx \dddot{\mu}_{\nu(i-1)} - \ddot{\mu}_{x(i)} \frac{\partial}{\partial \mu_{x(i)}} g(\mu_{x(i)}, \mu_{\nu(i)}) - \ddot{\mu}_{\nu(i)} \frac{\partial}{\partial \mu_{\nu(i)}} g(\mu_{x(i)}, \mu_{\nu(i)}) \\
\varepsilon_{x(i)} &= \dot{\mu}_{x(i)} - f(\mu_{x(i)}, \mu_{\nu(i)}) \\
\dot{\varepsilon}_{x(i)} &\approx \ddot{\mu}_{x(i)} - \dot{\mu}_{x(i)} \frac{\partial}{\partial \mu_{x(i)}} f(\mu_{x(i)}, \mu_{\nu(i)}) - \dot{\mu}_{\nu(i)} \frac{\partial}{\partial \mu_{\nu(i)}} f(\mu_{x(i)}, \mu_{\nu(i)}) \\
\ddot{\varepsilon}_{x(i)} &\approx \dddot{\mu}_{x(i)} - \ddot{\mu}_{x(i)} \frac{\partial}{\partial \mu_{x(i)}} f(\mu_{x(i)}, \mu_{\nu(i)}) - \ddot{\mu}_{\nu(i)} \frac{\partial}{\partial \mu_{\nu(i)}} f(\mu_{x(i)}, \mu_{\nu(i)}). \quad (66)
\end{aligned}
$$

Каузальные и скрытые состояния окружающего мира могут быть сложно организованы и иметь между собой зависимости, образуя своего рода иерархию и представляться аналогично порождающей модели. Однако в своих работах Фристон [1]-[12], [43], [44], [49], [64], [65], [123] не использует моделей с многоуровневыми каузальными и скрытыми состояниями среды, хотя такая возможность в его теории имеется. Тогда уравнения такого порождающего процесса аналогичны по виду уравнениям порождающей модели с той лишь разницей, что $\boldsymbol{\mu}_{x(i)}$, $\boldsymbol{\mu}_{v(i)}$, $\boldsymbol{\varepsilon}_{x(i)}$, $\boldsymbol{\varepsilon}_{\nu(i)}$ заменяются на $\boldsymbol{x}_i$, $\boldsymbol{v}_i$, $\boldsymbol{w}_i$, $\boldsymbol{z}_i$, соответственно.

Этот иерархический подход дополняется и расширяется идеей *предиктивного кодирования*. Ещё раз приведем её здесь. Идея *предиктивного кодирования* (в данном случае по отношению к зрительной системе) состоит в том, что *обратные связи от вышестоящих в иерархии областей зрительной коры несут прогнозы (предсказания) нейронной активности для нижестоящих областей зрительной коры, а прямые связи от нижестоящих областей зрительной коры возвращают остаточные ошибки между прогнозами (предсказаниями) и фактической активностью нижестоящих областей коры к вышестоящим областям зрительной коры* [30].

Мы можем сказать, что в теории Фристона это определение фактически приобретает такую форму: *обратные связи от вышестоящих в иерархии областей мозга несут прогнозы (предсказания) нейронной активности для нижестоящих областей мозга, а прямые связи от нижестоящих областей мозга возвращают остаточные ошибки между прогнозами (предсказаниями) и фактической активностью нижестоящих областей мозга к вышестоящим областям мозга.*

В контексте этого определения предиктивного кодирования представленная выше схема, с обобщёнными каузальными и обобщёнными скрытыми состояниями, составляющими на каждом уровне иерархии модуль

состояний, должна быть модифицирована. К этой схеме Фристон добавляет новый модуль, который мы будем называть *модулем ошибок предсказания* [1]-[3], [5]. В результате этой модификации состояния каждого отдела мозга описываются, во-первых, обобщённым каузальным состоянием, во-вторых, обобщённым скрытым состоянием (вместе составляющими модуль состояний), и в-третьих, модулем ошибок предсказания.

В соответствии с этим, согласно Фристону [3], [93], предполагается следующее:

- во-первых, *наличие популяции нейронов, чья активность кодирует условное математическое ожидание или условную моду модулей состояний, составляющих ненаблюдаемые и наблюдаемые состояния разных уровней иерархии в мозгу*;

- во-вторых, *наличие популяции нейронов, чья активность кодирует ошибки предсказания каждого состояния*;

- в-третьих, *модули ошибок предсказания получают сообщения от модулей состояний того же уровня иерархии и уровня выше*;

- в-четвертых, *модули состояния получают сообщения от модулей ошибок того же уровня иерархии и уровня ниже.*

С модулями ошибок предсказания Фристон [3] связывает ещё одну величину, которую он определяет следующим образом:

$$\begin{aligned}
\xi_{\nu(i)} &= \tilde{C}_z^{-1}\varepsilon_{\nu(i)} \\
\xi_{x(i)} &= \tilde{C}_w^{-1}\varepsilon_{x(i)} \\
\xi_{\theta(i)} &= \tilde{C}_\theta^{-1}\varepsilon_{\theta(i)} \\
\xi_{\gamma(i)} &= \tilde{C}_\gamma^{-1}\varepsilon_{\gamma(i)}.
\end{aligned} \tag{67}$$

Величина ξ — это средневзвешенная по точности обобщённая ошибка предсказания.

Здесь следует сказать ещё об одном важном понятии, фигурирующем в теории Фристона [1]-[12], [43], [44], [49], [64], [65], [123] — это «достаточная статистика». Так как в качестве одной из важнейших гипотез в теории Фристона является приближение Лапласа, состоящее в том, что все вероятностные распределения имеют форму гауссиана, то для того чтобы знать всё об этих распределениях, достаточно лишь задать математическое ожидание и матрицу ковариации. Поэтому, например, нет никакой необходимости полностью передавать всё вероятностное априорное распределение из одного слоя иерархической модели в другой (из

вышележащего в нижележащий), а можно передавать лишь достаточную статистику. Эта полученная нижележащим слоем из вышележащего достаточная статистика содержит всю информацию о форме априорного распределения. Аналогично и для передачи сообщений вверх по иерархии: вместо того, чтобы передавать весь набор ошибок по всему распределению, вверх по иерархии передается только средневзвешенная ошибка.

Теперь, когда все ключевые понятия введены, а предположения сделаны, можно, наконец, переходить к основной системе дифференциальных уравнений теории Фристона.

5. Гипотеза о градиентном спуске по информационной свободной энергии

Для того чтобы завершить построение математического формализма теории Фристона требуется найти уравнение, соответствующее такой динамике условного математического ожидания, при которой ИСЭ могла бы минимизироваться. Одним из простейших таких уравнений является дифференциальное уравнение, соответствующее методу градиентного спуска (Градиентный спуск — метод нахождения локального минимума или максимума целевой функции с помощью движения вдоль градиента.). Такое предположение о динамике математического ожидания и составляет *гипотезу о градиентном спуске по информационной свободной энергии (ИСЭ)* [1]-[12]. В этом случае динамика условного математического ожидания представляется, в сущности, как градиентный спуск по свободной энергии, то есть

$$\begin{aligned}\dot{\boldsymbol{\mu}}_{\nu(i)} &= -\frac{\partial}{\partial \boldsymbol{\mu}_{\nu(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) \\ \dot{\boldsymbol{\mu}}_{x(i)} &= -\frac{\partial}{\partial \boldsymbol{\mu}_{x(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)})\end{aligned} \tag{68}$$

Правая часть этих уравнений, как нетрудно заметить, является антиградиентом (то есть градиент со знаком минус) от ИСЭ. Соответственно, правая часть уравнения 51 должна быть продифференцирована по векторам, представленным в уравнениях, представленных выше.

Помимо этого Фристон [2], [7] вводит требование, чтобы *путь условной моды (математического ожидания) был равен моде пути, когда ИСЭ равна нулю*, то есть

$$\dot{\boldsymbol{\mu}} = D\boldsymbol{\mu}. \tag{69}$$

Обоим этим требованиям удовлетворяет следующая система дифференциальных уравнений(Более подробно по поводу того, как Фристон при-

ходит к представленному уравнению, можно познакомиться в следующей работе [2].):

$$\dot{\boldsymbol{\mu}}_{\nu(i)} = D\boldsymbol{\mu}_{\nu(i)} - \frac{\partial}{\partial \boldsymbol{\mu}_{\nu(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)})$$

$$\dot{\boldsymbol{\mu}}_{x(i)} = D\boldsymbol{\mu}_{x(i)} - \frac{\partial}{\partial \boldsymbol{\mu}_{x(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}). \tag{70}$$

Уравнение 89 задает траекторию изменения условного математического ожидания, в соответствии с которой минимизируется ИСЭ.

Фристон [2], [7] дополняет эту систему уравнений ещё несколькими уравнениями, отражающими изменение сенсорных состояний $\boldsymbol{s}$, скрытых состояний среды $\boldsymbol{x}$ и их причин $\boldsymbol{\nu}$, шумов скрытых состояний $\boldsymbol{z}$ и их причин $\boldsymbol{w}$ и априорных вероятностей наивысшего уровня иерархии $\boldsymbol{\eta}$ (то есть порождающий процесс в обобщённых координатах). А также он добавляет сюда дифференциальные уравнения, соответствующие динамике обучения и действия. В итоге имеем следующую систему дифференциальных уравнений:

$$\begin{aligned}
&\boldsymbol{s} = g(\boldsymbol{x}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{a}) + \boldsymbol{z} \\
&\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{\nu}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{a}) + \boldsymbol{w} \\
&\dot{\boldsymbol{\nu}} = D\boldsymbol{\nu} \\
&\dot{\boldsymbol{z}} = D\boldsymbol{z} \\
&\dot{\boldsymbol{w}} = D\boldsymbol{w} \\
&\dot{\boldsymbol{\eta}} = D\boldsymbol{\eta} \\
&\dot{\boldsymbol{\mu}}_{\nu(i)} = D\boldsymbol{\mu}_{\nu(i)} - \frac{\partial}{\partial \boldsymbol{\mu}_{\nu(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) \\
&\dot{\boldsymbol{\mu}}_{x(i)} = D\boldsymbol{\mu}_{x(i)} - \frac{\partial}{\partial \boldsymbol{\mu}_{x(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) \\
&\dot{\boldsymbol{\mu}}_{\theta(i)} = D\boldsymbol{\mu}_{\theta(i)} - \frac{\partial}{\partial \boldsymbol{\mu}_{\theta(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) \\
&\dot{\boldsymbol{\mu}}_{\gamma(i)} = D\boldsymbol{\mu}_{\gamma(i)} - \frac{\partial}{\partial \boldsymbol{\mu}_{\gamma(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) \\
&\ddot{\boldsymbol{\mu}}_{\theta(i)} = -\frac{\partial}{\partial \boldsymbol{\mu}_{\theta(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) - k\dot{\boldsymbol{\mu}}_{\theta(i)} \\
&\ddot{\boldsymbol{\mu}}_{\gamma(i)} = -\frac{\partial}{\partial \boldsymbol{\mu}_{\gamma(i)}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)}) - \boldsymbol{\mu}_{\gamma(i)} \\
&\dot{\boldsymbol{a}} = -\frac{\partial}{\partial \boldsymbol{a}} F(\boldsymbol{s}, \boldsymbol{\mu}_{x(i)}, \boldsymbol{\mu}_{\nu(i)}, \boldsymbol{\mu}_{\theta(i)}, \boldsymbol{\mu}_{\gamma(i)})
\end{aligned} \tag{71}$$

где k – это скорость обучения, $\boldsymbol{a}$ – это действие, все остальные обозначения остаются теми же, как описано ранее.

У Фристона нет какого-то конкретного названия для данной системы дифференциальных уравнений. К тому же вид этой системы, а также количество содержащихся в ней уравнений, несколько видоизменяется от статьи к статье. Поэтому мы назовём эту систему дифференциальных уравнений «системой уравнений Фристона».

Шум скрытых и каузальных состояний среды задается как аналитическая функция [89]. В качестве такой функции используется гауссиан с математическим ожиданием равным нулю [89].

Таким образом, принятием *гипотезы о градиентном спуске по ИСЭ*, завершается построение математического формализма теории Фристона в непрерывной формулировке. В итоге мы, вслед за Фристоном, приходим к системе уравнений 71, содержащей, во-первых, порождающий процесс как уравнения динамики скрытых, каузальных состояний среды и их шумов, во-вторых, уравнения динамики гипотез о математическом ожидании скрытых и каузальных состояний, представляющие из себя, в сущности, обратную модель по отношению к порождающему процессу, и позволяющие предсказывать динамику порождающего процесса, в-третьих, уравнения динамики переменных, связанных с обучением и действием.

В следующем разделе мы рассмотрим алгоритм, при котором реализуется схема минимизации ИСЭ и численно решается система уравнений Фристона.

6. Динамическая максимизация математического ожидания

Для реализации схемы минимизации ИСЭ Фристон [2] предлагает процедуру (алгоритм), называемую динамической максимизацией математического ожидания (*Dynamical Expectation Maximization*).

Динамическая максимизация математического ожидания — это процедура (алгоритм) минимизации ИСЭ, являющаяся модификацией EM алгоритма (алгоритма максимизации математического ожидания) и содержащая в себе численное решение системы уравнений Фристона.

Эта процедура состоит из 3 шагов: D-шаг, E-шаг и M-шаг.

На D-шаге производится численное решение уравнений 1-7 в системе уравнений 71. Для численного решения этого уравнения применяется метод Ozaki (1992) [94].

Согласно методу Ozaki (1992) система уравнений Фристона приближается соотношением в общем виде, предстающем как:

$$\boldsymbol{\mu}_{x(i)[n+1]} = \boldsymbol{\mu}_{x(i)[n]} + (e^{J\Delta t} - I)J^{-1}\frac{d}{dt}\boldsymbol{\mu}_{x(i)}$$
$$\boldsymbol{\mu}_{\nu(i)[n+1]} = \boldsymbol{\mu}_{\nu(i)[n]} + (e^{J\Delta t} - I)J^{-1}\frac{d}{dt}\boldsymbol{\mu}_{\nu(i)}, \tag{72}$$

где J — якобиан, I — единичная матрица, Δt - временной шаг.

Для вычисления якобиана J нужно взять вариационные производные поочередно по $\boldsymbol{s}$, $\boldsymbol{\mu}_x$ и $\boldsymbol{\mu}_\nu$. Получающаяся матрица очень велика по размеру, поэтому в тексте нашей статьи мы её не приводим. Читателям желающим более подробно ознакомится с элементами этой матрицы, мы рекомендуем обратиться к следующим работам [2], [4].

На E-шаге уравнения Фристона приближаются следующим соотношением:

$$\boldsymbol{\mu}_{\theta(i)[n+1]} = \boldsymbol{\mu}_{\theta(i)[n]} + J^{-1}\frac{d}{dt}\boldsymbol{\mu}_{\theta(i)}. \tag{73}$$

На M-шаге уравнения Фристона приближаются следующим соотношением:

$$\boldsymbol{\mu}_{\gamma(i)[n+1]} = \boldsymbol{\mu}_{\gamma(i)[n]} + J^{-1}\frac{d}{dt}\boldsymbol{\mu}_{\gamma(i)}. \tag{74}$$

На E- и M-шагах, как легко заметить, Фристон делает значительное упрощение выражений, благодаря которому удается избежать расчёта матричной экспоненты, что значительно облегчает вычисления [2].

За более подробной информацией, касающейся расчёта вышеприведенных величин, рекомендуем обратиться к программному коду, доступному на сайте, посвященном *Statistical Parametric Mapping (SPM)* [89].

Теперь, когда все основные понятия теории Фристона введены и математический аппарат полностью сформулирован, можно, наконец, перейти к непосредственным результатам, объясняющим разные аспекты работы мозга, которые можно получить, применяя эту теорию, а также обсудить эти результаты и само содержание теории.

7. Модельные эксперименты из работ Фристона

Из примерно шести десятков моделей, созданных Фристоном и его командой за почти 2 десятилетия их работы, с его теорией в непрерывной формулировке связаны примерно половина из них, а уже из этой половины с нейрональными и мозговыми процессами — чуть больше десятка.

Интересно, что несмотря на то, что Фристон позиционирует свою теорию чуть ли не как универсальную для объяснения самых разнообразных биологических систем и происходящих в них явлений, среди

собственных моделей Фристона очень много таких, которые, по нашему мнению, совершенно никак не связаны с объяснением каких бы то ни было биологических явлений, поскольку относятся к другим областям знаний. Наиболее яркими представителями таких моделей являются модели, связанные с так называемой «mountain car problem», встречающиеся как минимум в 3 работах [1], [6], [44]. Это проблема скорее из области теории оптимального управления, а не биологии.

Мы подробно разберём модели из двух работ. Как нам представляется, эти модели самые показательные, в плане демонстрации того, как применяется теория Фристона самим автором, а также как в этих моделях проявляются некоторые из основных проблем и недостатков теории Фристона.

В работе «*Active Inference and Learning in the Cerebellum*» [9] представлена модель, способная, согласно заявлениям авторов, воспроизводить основные особенности поведения, характерные для отставленного (delay conditioning) и следового (trace conditioning) мигательного условного рефлексов, спонтанного моргания и реакции испуга (правда, понять из текста, каким именно образом была промоделирована реакция испуга, представляется крайне затруднительным). Порождающая модель имеет 2 иерархических уровня. Скрытые состояния на 1 уровне связаны с гипотезами об условных и безусловных раздражителях и морганиях. На этом уровне проприоцептивные предсказания (моторные команды) представляют собой гипотезы о математических ожиданиях о спонтанных морганиях и безусловных стимулах. Скрытые состояния на втором уровне воплощают ощущение времени через динамику уравнения аналогичного уравнению Лотки-Вольтерры. Фристон связывает эту динамику с работой центрального генератора паттерна, который посещает последовательность неустойчивых фиксированных точек в определенной последовательности (то есть, гетероклинический цикл). Гетероклинический цикл (гетероклинический цикл — это топологический круг точек равновесия и соединяющих гетероклинических орбит) генерирует последовательность скрытых состояний. Эти состояния передаются на первый уровень через функцию *softmax* и генерируют спонтанные моргания или предсказывают условные раздражители. Гетероклинический цикл имеет последовательность трех состояний, которые приводят к циклу. Цикл сам по себе генерирует спонтанные моргания, первое состояние последовательности генерирует условный стимул. Последующие нестабильные фиксированные точки играют роль следа для эхо-состояний, которые позволяют обучаться или связывать последующие безусловные раздражители, которые следуют за условными раздражителями через некоторое время. Безусловный стимул моделируется как гауссова функция.

Обучение, согласно Фристону, соответствует хеббовской пластичности, которая сводит к минимуму ошибку предсказания в связях между математическими ожиданиями в отношении безусловного раздражителя и математическими ожиданиями первого уровня в отношении безусловного раздражителя для отставленного условного рефлекса и вторым уровнем эхо-состояний для следового условного рефлекса. Фристон ещё пишет о том, что «обучение при отставленном мигательном условном рефлексе заключается в том, что преобладают изменения в связи между математическими ожиданиями относительно условного раздражителя (кодируемого клетками Пуркинье в коре мозжечка) и апостериорными предсказаниями о безусловном раздражителе (в промежуточном ядре)» и «при отставленном условном рефлексе именно связь с третьей скрытой причиной опосредует этот условный рефлекс». «Это происходит потому, что задержка означает, что эта скрытая причина активна в то время, когда вызывается безусловный ответ. Важно отметить, что это означает, что задействован другой набор связей, а именно: связи между ядрами моста и промежуточным ядром». Порождающая модель состоит из следующих уравнений:

$$
\begin{aligned}
f^{(2)} &= \theta_x^{(2)} + \frac{1}{8}x^{(2)} + I \approx \dot{x}^{(2)} \\
g^{(2)} &= \sigma(x^{(2)}) = \nu^{(1)} \\
f^{(1)} &= \begin{pmatrix} f_{CS}^{(1)} \\ f_{US}^{(1)} \\ f_{EB}^{(1)} \end{pmatrix} = \begin{pmatrix} \nu_1^{(1)} - x_{CS}^{(1)} \\ \theta_\nu^{(1)}\nu_1^{(1)} + \theta_x^{(1)}x_{CS}^{(1)} - x_{US}^{(1)} \\ \nu_6^{(1)} - x_{EB}^{(1)} \end{pmatrix} \approx \dot{x}^{(1)} \\
g^{(1)} &= \begin{pmatrix} g_{CS}^{(1)} \\ g_{US}^{(1)} \\ g_{EB}^{(1)} \end{pmatrix} = \begin{pmatrix} x_{CS}^{(1)} \\ x_{US}^{(1)} \\ x_{US}^{(1)} + x_{EB}^{(1)} \end{pmatrix} \approx s.
\end{aligned}
\tag{75}
$$

Соответственно, для того, чтобы получить систему уравнений, которая используется Фристоном для моделирования в этой статье, нужно взять уравнение 90 и подставить вместо $f^{(1)}$, $f^{(2)}$, $g^{(1)}$ и $g^{(2)}$, фигурирующих в уравнении 90 как содержимое векторов $\tilde{f}$ и $\tilde{g}$, правые части уравнения 94. Недостающие переменные можно найти в файле $ADEM\_eyeblink.m$ из пакета SPM и в тексте обсуждаемой статьи.

Экстероцептивные ошибки предсказания Фристон в своей модели для безусловного, условного рефлексов и спонтанного моргания связывает с ядрами моста, проприоцептивные ошибки предсказания — с красным ядром, соматосенсорные ошибки предсказания — с ядрами нижней оливы, математические ожидания о скрытых состояниях (и их ошибках), опосредующих спонтанные моргания — с ядрами моста, математические ожидания о скрытых состояниях (и их ошибках), опосредующих

безусловные раздражители — с промежуточным ядром, математические ожидания о скрытых состояниях (и их ошибках), опосредующих условные раздражители — с корой мозжечка, скрытые состояния на втором уровне иерархии — с бледным шаром и гиппокампом или миндалина, связанные (второй уровень) скрытые причины — с ядром моста.

Фристон пишет о том, что схема предиктивного кодирования для спонтанного моргания, безусловного мигательного рефлекса, отставленного и следового мигательного условных рефлексов анатомически релевантна. Однако нетрудно заметить, что она имеет некоторые особенности, никем не описанные, да и, насколько нам известно, просто неверные.

Во-первых, в схеме Фристона имеется прямая связь между корой мозжечка и ядрами моста. Действительно, известно, что часть аксонов клеток Пуркинье заканчивается на вестибулярных ядрах ствола мозга (мозжечково-вестибулярный путь) [95]. Из этих ядер только верхнее вестибулярное ядро расположено в мосте, а остальные, то есть медиальное, латеральное и нижнее вестибулярные ядра расположены в продолговатом мозге [96]. Однако нами не было обнаружено ни одного упоминания в литературе об участии вестибулярных ядер ни в безусловном, ни в условном отставленном и следовом мигательных рефлексах, ни в спонтанном моргании. Таких данных не приводит и Фристон в своей статье. Эту связь Фристон использует при объяснении следового мигательного условного рефлекса. Исходя из написанного, релевантность всех тех толкований и выводов, которые делает Фристон с использованием этой связи в отношении следового мигательного условного рефлекса представляется сомнительной.

Во-вторых, согласно Фристону самопроизвольное (спонтанное) моргание порождается гетероклиническим циклом скрытых состояний и связанных с ними скрытых причин и опосредуется нисходящими предсказаниями от центра моргания в бледном шаре до ядер моста и красного ядра, то есть речь идёт о пути: бледный шар и гиппокамп (Фристон здесь не поясняет, почему бледный шар и гиппокамп рассматриваются вместе, будто это единое образование) $\rightarrow$ ядра моста $\rightarrow$ красное ядро. Однако, по нашему мнению, путь ядро моста $\rightarrow$ красное ядро для спонтанного моргания сомнителен, также как и путь бледный шар-гиппокамп $\rightarrow$ ядра моста. Нам не удалось найти в литературе ни одного источника подтверждающего участие этих путей в спонтанном моргании. Путь, который нам удалось найти в литературе [97], [98]: бледный шар $\rightarrow$ ретикулярная часть чёрной субстанции $\rightarrow$ верхнее двухолмие $\rightarrow$ большое ядро шва $\rightarrow$ ядро тройничного нерва $\rightarrow$лицевое ядро. Соответственно и всё то, что написано у Фристона о спонтанных морганиях, вызывает вопросы.

В-третьих, Фристон пишет о том, что «самопроизвольное моргание опосредуется нисходящими предсказаниями от центра моргания в блед-

ном шаре к ядрам моста и, наконец, к красному ядру». Другими словами, путь, представленный Фристоном, достаточно хорошо вписывается в его интерпретацию предиктивного кодирования. Хотя и тут есть вопросы следующего плана. Известно, что проекции из бледного шара — это тормозные проекции, проекции из ядер моста к красному ядру — или тормозные, или возбуждающие [99]. То есть даже в пути, предложенным Фристоном, имеется никак не описанная возможность того, что от бледного шара к ядрам моста предсказание движется по *тормозным* путям, а от ядер моста к красному ядру по *возбуждающим*. Каких-то разъяснений по поводу подобной ситуации в работах Фристона нам найти не удалось.

Кроме того, путь для спонтанного моргания найденный в нами в литературе [97], [98], опять же по нашему мнению, не вписывается во фристоновскую концепцию предиктивного кодирования. Дело в том, что помимо критики, связанной с возбуждающими и тормозными связями, где тормозные связи: бледный шар → ретикулярная часть чёрной субстанции, ретикулярная часть чёрной субстанции → верхнее двухолмие, а возбуждающие: верхнее двухолмие → большое ядро шва, ядро тройничного нерва → лицевое ядро, вызывает вопросы к теории Фристона ещё одна связь. Это связь большое ядро шва → ядро тройничного нерва, являющаяся серотонинергической, нейромодулирующей. Для нейромодулятора серотонина, так же как и для других нейромодуляторов [**?**], Фристон в нескольких своих работах [101], [102] предлагает играть роль в кодировании точности. Иначе говоря, цепочка передачи предсказаний, и соответственно ошибок предсказаний, характерная в целом для предиктивного кодирования и его фристоновской интерпретации в частности, разрывается связью, связанной с кодированием точности. Для такого случая в работах Фристона нами не было обнаружено аналогичных примеров и объяснений.

В-четвертых, при безусловном мигательном рефлексе, согласно Фристону, нисходящие проприоцептивные предсказания (или моторные команды) исходят из промежуточного ядра. При активном выводе безусловный рефлекс основывается на ожиданиях совместного возникновения безусловного стимула и реакции. Это ожидание вызывается безусловным стимулом и впоследствии вызывает действие посредством нисходящих предсказаний ожидаемых проприоцептивных последствий. Соматосенсорные ошибки предсказания, согласно Фристону, связаны ещё и с путём: ядро тройничного нерва → нижняя олива → промежуточное ядро. Соматосенсорные ошибки предсказания вызывают безусловный ответ. Математическое ожидание относительно безусловного ответа генерируют нисходящие предсказания, вызывающие моргание. Математическое ожидание относительно безусловного ответа связано с путями:

промежуточное ядро → красное ядро и промежуточное ядро → нижняя олива. Моргание, вызванное нисходящими предсказаниями, отменяет соматосенсорную ошибку предсказания.

И опять же, путь и, соответственно, связанный с ним механизм, сомнителен. Согласно литературным данным [103], [104], для безусловного сигнала есть, во-первых, короткий путь через связь ядра тройничного нерва с лицевым ядром, то есть путь безусловного сигнала, связанный с R1-пиком реакции самопроизвольного моргания: глаз → ядро тройничного нерва → лицевое ядро → мышца глаза; во-вторых, есть другой путь, связанный с R2-пиком: глаз → каудальная часть ядра тройничного нерва → латеральное тегментальное поле (*lateral tegmental field*) продолговатого мозга → лицевое ядро → мышца глаза. Промежуточное ядро тормозится нейронами Пуркинье и нами не было обнаружено в литературе данных, подтверждающих его участие в мигательном безусловном рефлексе. Для безусловного рефлекса, говоря языком теории Фристона, исходя из связи: ядро тройничного нерва → лицевое ядро, получается, что ядро тройничного нерва передает ошибку предсказания на лицевое ядро, а лицевое ядро передает ошибку предсказания в качестве моторной команды, что не вписывается в предлагаемую Фристоном модель безусловного мигательного рефлекса. Кроме того, такой путь не вписывается в несколько характеристик фристоновской интерпретации предиктивного кодирования, предполагающих, что каждый блок математического ожидания связан реципрокно с блоком ошибок предсказания и что блок ошибок предсказания отправляет сигнал только в блок математического ожидания. Кроме того, учитывая вышеописанное и тот факт из теории Фристона, что «информация, передаваемая первичными сенсорными афферентами, становится сигналом ошибки предсказания только тогда, когда встречается предсказание» [72], возникает вопрос о том, откуда же тогда ядро тройничного нерва получает предсказания?

При выработке отставленного условного рефлекса безусловные и условные раздражители сосуществуют во времени, но условный раздражитель начинается раньше безусловного. Поэтому достаточно логично ожидать сопряженных по времени экстероцептивных и проприоцептивных сигналов моргания глаз. Это требует, согласно Фристону, ожиданий условных стимулов, кодируемых клетками Пуркинье при получении восходящих ошибок предсказания от ядер моста (через моховидные и параллельные волокна) и ожиданий относительно безусловного стимула от промежуточного ядра и ядра нижней оливы (через лазящие волокна). После того, как эти ожидания были индуцированы условными и безусловными стимулами, они обеспечивают нисходящие проприоцептивные предсказания к промежуточным ядрам и corollary разряду вдоль параллельных волокон.

Короче говоря, отставленный мигательный условный рефлекс включает реципрокные связи между промежуточным ядром и корой мозжечка.

Схемам для отставленного мигательного условного рефлекса, приводимым в литературе [105]-[107], фристоновская схема этого рефлекса в целом соответствует.

В литературе по отставленному мигательному условному рефлексу, несмотря на то, что так называемые «минимальные» нейроанатомические схемы, лежащие в его основе, хорошо описаны, накопилось огромное количество трудно сопрягаемой друг с другом информации, так что физиология отставленного мигательного условного рефлекса ещё очень далека до понимания. Например, до сих пор представляется неясным, каким образом развивается реакция паузы в импульсации клетки Пуркинье в ответ на условный стимул [108], и как сопрягаются между собой возможные объяснения этой паузы, приводимые в литературе [109]-[111], какова роль промежуточных ядер в мигательном условном рефлексе, особенно с учетом работ Delgado-Garsia и Gruart [112]-[114] и работ, связанных с действием гармалина на глубокие ядра мозжечка [99], [115], [116], причину того, почему у разных видов животных (например, крыс и кроликов) значимость вклада миндалины в выработку отставленного условного рефлекса существенно различается [117], [118], роль гиппокампа в ассоциации условного и безусловного стимулов [**?**] и т.д.

Нам представляется сомнительным, что в плане понимания такой физиологии поможет предлагаемая модель Фристона, по крайней мере в той форме, в которой она сформулирована к настоящему времени.

На наш взгляд, важной нерешённой проблемой теории Фристона является то, что она не содержит метода по переводу, перекодированию тезисов и результатов, сформулированных на языке теории Фристона (то есть тезисов и результатов, где используются специальные термины, такие как «предсказания», «ошибки предсказания», «скрытые состояния», «причины» и т. д.) на профессиональный язык нейрофизиологии, в котором, в применении к мигательному условному рефлексу, используются такие термины как: паттерн импульсации нейронов Пуркинье и нейронов промежуточного ядра, пластических перестроек и биохимических каскадов, сопровождающих формирование отставленного мигательного условного рефлекса и т. д. И в целом, возникает вопрос, каким образом подход, предлагаемый Фристоном, может дать хоть какую-то полезную информацию о том богатстве интегративных свойств нейрона, которое имеет место, исходя из анализа литературных данных [120]-[122].

Фристон неоднократно пишет в своих работах следующее: «*... Эти наблюдения согласуются с экспериментальными данными ...*», «*... смоделированные и эмпирические эффекты ... согласованы ...*» и «*... экспе-*

рименты с моделируемым повреждением ... на удивление хорошо воспроизводят эмпирические результаты ...». Однако в отношении того, какие именно имеются в виду экспериментальные данные, у него нигде не говорится. Поэтому к этим высказываниям сразу появляются вопросы: «С какими наблюдениями и экспериментами согласуются результаты моделирования? Каким образом? Как это количественно и качественно оценить?»

Активность в нейронных популяциях, кодирующих математические ожидания и ошибки предсказания, или, что то же самое, динамика скрытых состояний и причин, связывается Фристоном с функцией временем перистимула. У Фристона не говорится о том, что означает выражение «функция времени перистимула». По нашему мнению, речь здесь может идти о скорости нейрональной импульсации, привязанной к действию релеватного стимула, что подтвержается следующим тезисом из разбираемой статьи: «*...нейрональная скорость разрядки кодирует ожидаемое состояние мира...*» И опять же не приводится конкретных работ, в которых можно было бы сравнить результаты, полученные в разбираемой работе Фристона с конкретными экспериментальными данными, и количественно и качественно оценить добротность фристоновского моделирования.

Стоит заметить, что в разбираемой работе величина функции перистимула измеряется в каких-то неизвестных, неописанных условных единицах, что говорит о, своего рода, условности или «игрушечности» упомянутой модели.

В отношении того, как Фристон в разбираемой работе промоделировал дефицит при поражении одного из путей для спонтанного моргания, отставленного и следового мигательного условных рефлексов, учитывая написанное выше, критика аналогична вышеприведенному для спонтанного моргания, отставленного и следового мигательного условных рефлексов. Почему? Потому что он использовал, как мы показали выше, какую-то «собственную архитектуру» для описания процессов, связанных с морганием, не ссылающуюся на конкретные эксперименты.

Таким образом, в отношении моделирования, представленного в разбираемой работе, можно сделать несколько выводов.

Во-первых, нейроанатомические схемы для спонтанного моргания, следового мигательного условного рефлекса сомнительны. Соответственно, сомнительно и всё то, что Фристон пишет в применении к этим схемам. Схема для отставленного мигательного условного рефлекса слишком упрощена и не учитывает многих важных особенностей (например, связанных с ролями миндалины и гиппокампа) этого условного рефлекса.

Во-вторых, не приводится информации о том, какие экспериментальные данные, из каких работ моделируются в этой статье и каким эмпирическим эффектам из каких работ соответствуют результаты из данной публикации.

В работе «*Cortical circuits for perceptual inference*» [123] Фристон с помощью моделирования пытается подкрепить схему предиктивного кодирования, относящуюся к кортикальным процессам. Однако делает он это весьма специфическим образом. Фристон пишет, что «...примером, который мы используем, является пение птиц, а эмпирические измерения, на которые мы ориентируемся, — это локальный полевой потенциал (LFP) или вызванные (ERP) ответы, которые можно регистрировать неинвазивно». Далее Фристон пишет о том, что с локальными полевыми потенциалами и вызванными ответами связана активность поверхностных пирамидных клеток и далее, что «...это означает, что мы можем отнести эти сигналы к ошибке предсказания; потому что поверхностные пирамидные клетки являются источником восходящих сообщений в головном мозге».

Модель использованная в этой работе, в основных чертах совпадает с моделью из работы «*Predictive coding under the free-energy principle*» [123]. Порождающая модель из файлов SPM ($DEM\_demo\_song\_priors.m$, $DEM\_demo\_song_omission.m$) не совпадает с уравнениями, представленными в тексте самой статьи. При моделировании в этой работе у Фристона есть только порождающая модель, но нет порождающего процесса.

Порождающая модель для пения птиц состоит из двух аттракторов Лоренца, где аттрактор более высокого уровня передает два управляющих параметра аттрактору более низкого уровня, который, в свою очередь, передает два управляющих параметра «синтетическому сиринксу», один из этих параметров управляет частотой (от двух до пяти кГц), а другой управляет амплитудой, или громкостью. Далее Фристон пишет о том, что «*...параметры аттрактора Лоренца были выбраны таким образом, чтобы генерировать короткую последовательность щебетов каждую секунду или около того*». Динамика 2-го аттрактора Лоренца на порядок медленнее, чем у другого аттрактора Лоренца расположенного в иерархии ниже. Затем он пишет о том, что «*...состояния более медленного аттрактора вводились в качестве управляющих параметров (числа Рэлея и Прандтля) для управления динамикой, демонстрируемой более быстрым. Эта динамика может варьироваться от аттрактора с фиксированной точкой, где все состояния первого равны нулю; вплоть до квазипериодического и хаотического поведения, когда значение числа Рэлея превышает соответствующий порог (около двадцати четырех) и вызывает бифуркацию. Поскольку более высокие состояния развива-*

ются медленнее, они включают и выключают нижний аттрактор, создавая отдельные песни, где каждая песня состоит из серии отдельных щебетаний. Эта порождающая модель генерирует спонтанные последовательности песен».

Уравнения порождающей модели, приводимые в тексте исходной статьи:

$$f^{(1)} = \begin{pmatrix} 18x_2^{(2)} - 18x_1^{(2)} \\ 32x_1^{(2)} - 2x_3^{(2)}x_1^{(2)} - x_2^{(2)} \\ 2x_1^{(2)}x_2^{(2)} - \frac{8}{3}x_3^{(2)} \end{pmatrix}$$

$$g^{(1)} = \begin{pmatrix} x_2^{(1)} \\ x_2^{(1)} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

$$f^{(2)} = \begin{pmatrix} 18x_2^{(1)} - 18x_1^{(1)} \\ \nu_1^{(1)}x_1^{(1)} - 2x_3^{(1)}x_1^{(1)} - x_2^{(1)} \\ 2x_1^{(1)}x_2^{(1)} - \nu_2^{(1)}x_3^{(2)} \end{pmatrix}$$

$$g^{(1)} = \begin{pmatrix} x_2^{(2)} \\ x_2^{(2)} \end{pmatrix} = \begin{pmatrix} \nu_1^{(1)} \\ \nu_2^{(1)} \end{pmatrix} \tag{76}$$

Начальные значения переменных и другие параметры приведены в файлах $DEM\_demo\_song\_priors.m, DEM\_demo\_song\_omission.m$.

Затем Фристон пишет о том, что он сгенерировал с помощью представленной выше порождающей модели одну сонограмму, а после этого её инвертировал.

Другими словами, он на основе инвертированной сонограммы демонстрирует с помощью моделирования, основанного на математическом аппарате его теории, что достоверные скрытые состояния и каузальные состояния могут быть найдены или, что то же самое, восстановлены.

Подобные результаты не вызывают удивления, потому что в математический аппарат теории Фристона заложен метод оптимизации — *DEM* алгоритм, — который и позволяет находить неизвестные значения параметров порождающей модели по известным данным. При всём этом сами данные были порождены порождающей моделью, что само по себе, по нашему мнению, существенно облегчает такой поиск.

Фристон также пишет о том, что он своим моделированием пытается показать, «как автономная динамика (связанных аттракторов Лоренца) создает порождающие модели сенсорного входа, которые ведут себя почти так же, как настоящий мозг, при электрофизиологическом измерении». Что в данном случае имеет в виду Фристон — из текста статьи

понять крайне затруднительно. В единственном рисунке, где фигурируют локальные полевые потенциалы, подписи к рисунку таковы, что из текста невозможно понять, каким образом эти потенциалы связаны с моделью Фристона в этой статье. И в целом, как и в случае предыдущей статьи данного автора, к нему возникают аналогичные вопросы в отношении того, какие экспериментальные данные он пытается промоделировать в своей модели? Как язык предсказаний, ошибок предсказаний переводится на язык локальных полевых потенциалов и вызванных ответов? Как можно количественно и качественно сопоставить предсказания, ошибки предсказания и т. д. с величинами локальных полевых потенциалов и вызванных ответов из каких-то конкретных экспериментов, в которых такие потенциалы измерялись? Какое отношение имеет моделирование пения птиц к локальным полевым потенциалам и вызванным ответам? Почему используется такая странная модель для моделирования явлений, на первый взгляд не имеющих друг к другу отношения?

Точно такие же проблемы, как в двух представленных работах, имеются и в работе «*Action and behavior: a free-energy formulation*», где представлено несколько моделей, одна из которых связана с окуломоторным контролем, другая — с сенсомоторной интеграцией, третья — с целенаправленным поведением.

Как написал Фристон, говоря про модель представленную в вышеразобранной статье [123] «*...здесь используется игрушечная модель*». И действительно, лучшего обозначения, по нашему мнению, в отношении не только этой модели, а практически всех изученных нами моделей из работ Фристона придумать трудно. Во всех моделях, которые нам попадались были одни и те же повторяющиеся из одной статьи в другую недостатки.

Во-первых, нигде Фристон не моделирует какие-то конкретные экспериментальные данные. Соответственно, количественно и качественно оценить, что промоделировал Фристон, сопоставив результаты его моделирования с конкретным экспериментом, не представляется возможным.

Во-вторых, анатомические основания представленных моделей вызывают вопросы. Откуда автор брал именно такую архитектонику связей? Каковы основания считать, что именно такая архитектоника связей, а не какая-либо другая, участвует в рассматриваемом явлении?

Все эти недостатки, присущие модельным экспериментам Фристона, представленные нами выше, по нашему мнению, говорят как минимум об «игрушечности», оторванности представленных моделей от нейрофизиологических экспериментов, а значит вполне возможно, что и об «игрушечности», оторванности от эксперимента и всей теории Фристона в целом.

8. Обсуждение

В предыдущих главах нами была изложена теория Фристона в непрерывной формулировке. При изложении этой теории мы старались освещать её сущность и содержание достаточно подробно и в понятной форме для неспециалистов по байесовской теории вероятностей, по ходу изложения обсуждая и критикуя содержание и вводимые положения этой теории.

Теория Фристона, являясь вероятностной (байесовской) теорией, по сравнению со всеми остальными объяснениями работы мозга, принятыми в нейрофизиологии, стоит особняком. Это легко можно понять по специфической терминологии и внутренней логике теории. Так, отталкиваясь от, в сущности, «кантианских» идей Гельмгольца о восприятии и продолжая, в отношении этих идей, «статистическую», «теоретико-информационную» линию Хинтона и соавторов, Фристон объединил под началом принципа (информационной) свободной энергии две гипотезы, касающиеся работы мозга — гипотезу байесовского мозга (кодирования) и гипотезу предиктивного кодирования, в собственной формулировке и с собственной интерпретацией. В дополнение ко всему этому, Фристон предложил ещё такие важные для его теории вспомогательные предположения, как: приближение Лапласа, гипотеза о математическом ожидании ненаблюдаемых состояний, предположение о локальной линейности, гипотеза о градиентном спуске по информационной свободной энергии. Принцип информационной свободной энергии, гипотеза байесовского мозга, гипотеза предиктивного кодирования, а также вышеупомянутые вспомогательные предположения и составляют здание теории.

Пройдемся кратко по всему изложенному выше.

В первом разделе, нами, на основе работ Фристона, рассматривается и обосновывается предлагаемый им принцип (информационной) свободной энергии. Для принципа информационной свободной энергии, являющегося базисом теории, Фристон в своих работах предлагает 2 варианта обоснования.

В первом варианте он исходит из гомеостаза и термодинамики, вводя основные понятия своей теории, такие как информационная энтропия и информационная свободная энергия, и делая при этом, как мы считаем, слабо обоснованный перескок в своих рассуждениях из термодинамики в теорию информации.

Во втором варианте он исходит из представления о том, что биологические системы — это эргодические случайные динамические системы (аттракторы). При таком варианте его подход получает серьезные математические обоснования, однако, как мы отметили, само представление о биологических системах как о эргодических случайных динамических аттракторах, требует доказательств.

Тесно связано с информационной свободной энергией такое понятие как информационное свободное действие, представляющее из себя интеграл по информационной свободной энергии по времени. Для этого понятия Фристон не приводит подробных разъяснений, однако введение этого понятия влечет за собой следствие — теория Фристона формулируется в форме обобщенных координат. Форма теории в виде обобщенных координат широко использует в физике, например, аналогичным образом сформулированы лагранжева и гамильтонова механика.

Во втором разделе нами, опять же на основании работ Фристона, рассматривается гипотеза байесовского мозга. В своих работах, вводя гипотезу байесовского мозга в теорию, Фристон начинает с рассуждений «кантианского» характера, разделяя мир и мозг стеной, причем информацию о мире мозг получает через посредника — сенсорные состояния. Из этих рассуждений делается вывод, что мозг может судить о состоянии мира только в вероятностном ключе, выдвигая гипотезы о том, в каком состоянии находится мир. Основой же для этих гипотез являются сенсорные (наблюдаемые) состояния, являющиеся зависимыми от состояний мира (ненаблюдаемых). Затем вводится понятие условной байесовской вероятности, позволяющее работать с наблюдаемыми и ненаблюдаемыми переменными, делая выводы о ненаблюдаемых переменных на основе наблюдаемых. Для расчёта такой условной вероятности Фристоном применяется теорема (формула) Байеса. Однако так как ненаблюдаемые состояния являются принципиально недоступными для мозга, то возникают сложности с таким расчетом. Большая часть всего последующего изложения состоит в преодолении этих сложностей путем манипуляций с формулой Байеса и введения дополнительных предположений. Проведя несколько манипуляций с формулой Байеса, Фристон вводит первое предположение о том, что мозг должен кодировать плотность распознавания (вариационное распределение), то есть в формулу Байеса вводится ещё одна функция. После нескольких манипуляций с модифицированной с помощью этой функции формулой Байеса, мы, вслед за Фристоном, приходим к уравнению, слагающемуся из информационной свободной энергии, информационной энтропии и ещё одной величины, называемой Фристоном информационной энергией Гиббса. Таким путем устанавливается связь между теоремой Байеса, используемой для расчета вероятностей ненаблюдаемых состояний, с одной стороны, и принципом (информационной) свободной энергии, введенным в предыдущем разделе, с другой стороны. Оказывается, что чем меньше при расчете становится величина информационной свободной энергии, тем точнее приближение для ненаблюдаемых состояний. Далее мы переводим выведенные формулы в обобщенные координаты. Однако в полученных формулах в обобщенных координатах имеется проблема неизвестности ненаблюдаемых

состояний, и мы, вслед за Фристоном, вводим второе предположение о том, что мозг делает гипотезы о математическом ожидании ненаблюдаемых состояний. А затем делаем третье предположение, называемое приближением Лапласа, заключающиеся в том, что всё распределения, фигурирующие в теории, приближаются гауссианом. Затем, после очередной серии математических преобразований, выяснилось, что для того чтобы рассчитать ИСЭ — достаточно задать только порождающую модель или, что то же самое, плотность совместного распределения сенсорных данных и гипотез о математическом ожидании ненаблюдаемых состояний (порождающей модели). Такая форма ИСЭ позволяет избавиться в формулах от неизвестных ненаблюдаемых состояний и оперировать только известными и доступными величинами: сенсорными данными и гипотезами о математическом ожидании ненаблюдаемых состояний. После этого мы показываем, что гипотезы о математическом ожидании ненаблюдаемых состояний состоят из гипотез о математическом ожидании скрытых состояний, причин, параметров и гиперпараметров. Затем мы расписываем функции других распределений, фигурирующих в формулах Фристона, а также демонстрируем внешний вид матриц ковариации. Затем вводится важная величина в теории Фристона называемая ошибкой предсказания. После этого мы показываем, какова динамика скрытых состояний и причин и то, из чего слагаются сенсорные состояния, в том числе в форме обобщенных координат.

Третий раздел посвящен включению в теорию Фристона ещё одного важного элемента — гипотезы предиктивного кодирования. В этом разделе рассуждения из второго раздела применяются уже к отделам (иерархическим структурам) мозга, в результате чего формализм теории Фристона обобщается на случай иерархической формы, не претерпевая при этом изменения общей формы уравнений теории. Благодаря этому происходит синтез принципа свободной энергии и гипотезы байесовского мозга, с одной стороны, и гипотезы предиктивного кодирования, с другой стороны. В результате такого синтеза передача информации в иерархии мозга теперь состоит в передаче вверх по иерархии ошибок предсказания о гипотезах о математическом ожидании ненаблюдаемых состояний и передачи вниз по иерархии предсказаний о математическом ожидании ненаблюдаемых состояний.

В четвёртом разделе завершается построение математического формализма теории Фристона. В этом разделе вводится ещё одно предположение — гипотеза о градиентном спуске по информационной свободной энергии. Эта гипотеза говорит о том, что Фристоном предлагается дифференциальное уравнение, соответствующее методу градиентного спуска, при котором динамика математического ожидания ненаблюдаемых состояний такова, что информационная свободная энергия имеет

тенденцию минимизироваться. Оканчивается этот раздел выводом системы уравнений Фристона.

В пятом разделе кратко излагается сущность алгоритма, применяемого Фристоном для решения системы уравнений Фристона.

Шестой раздел посвящен разбору нескольких моделей из работ Фристона, в которых его теория применяется на практике *in silico*. В результате анализа этих работ обнаруживаются серьезные проблемы, присущие как теории Фристона в целом, так и частные недостатки, характерные для рассмотренных работ.

Итак, изложив и проанализировав теорию Фристона, а также в целом все те работы Фристона и некоторых других авторов, на которые мы опирались, мы приводим ниже тезисно те достоинства и недостатки, которыми обладает, по нашему мнению, представленная теория.

Достоинства теории Фристона

1) *Универсальность, приложимость к широкому кругу нейрональных явлений.* При обсуждении модельных экспериментов из работ Фристона, видно, что с помощью своей теории Фристон моделирует такой широкий спектр явлений как разные виды мигательных условных рефлексов, мигательный безусловный рефлекс, спонтанное моргание, локальный полевой потенциал (local field potential) и вызванные потенциалы, окуломоторный контроль, сенсомоторную интеграцию, целенаправленное поведение и т. д. Другое дело, каким образом это всё моделируется. В этом смысле стоит заметить, что приложимость следует отличать от применимости. Применимость может сильно отставать от приложимости.

2) *Гибкость, возможность видоизменения формы порождающей модели в зависимости от того, что мы хотим описывать.* Теория Фристона позволяет создавать в рамках принятого формализма порождающие модели с произвольным количеством уровней и произвольной формы. Причем общая форма системы уравнений Фристона сохраняется, и не требуется её переделка под каждый конкретный случай. Гибкость теории Фристона подчеркивает и программное обеспечение SPM, где модели из разных работ Фристона отличаются, главным образом, только уравнениями для порождающего процесса и порождающей модели и некоторыми вспомогательными величинами. Другими словами, не требуется переписывать всю систему уравнений Фристона под каждый конкретный случай, под каждое новое моделирование.

3) *Легкость расширения теории, если такое расширение связано с введением векторных величин.* Теория Фристона исходно использовалась с урезанным количеством векторных величин по сравнению

с системой уравнений Фристона в её нынешней форме. Например, в ранних работах Фристона [2], [3], [5] не использовалось такое понятие как действие (*action*) (не путать со свободным действием), а в более поздних работах это понятие было введено [4], [6], причем существенного пересмотра системы уравнений Фристона это нововведение не потребовало.

Недостатки теории Фристона и вопросы к ней

1) *Кантианство, неявно лежащее в основе теории и, соответственно, связанная с ней критика кантианства.* Для начала покажем, почему теория Фристона — это кантианская теория. Помимо кантианских корней теории Фристона и других родственных теорий [16], [17], обнаруженных Л. Свонсоном [18] (за подробностями отсылаем читателя к упомянутой работ), а также других работ [124]-[127], указывающих на кантианство этой теории, можно привести в поддержку кантианства теории Фристона следующую аргументацию. Как мы писали выше, вся теория, в сущности, вращается вокруг формулы Байеса. А в формуле Баейса содержатся вероятности, связанные с наблюдаемыми и ненаблюдаемыми состояниями. Соответственно можно сказать, что вся теория вращается вокруг ненаблюдаемых и наблюдаемых состояний. Ненаблюдаемые состояния называют так потому, что они не являются непосредственно доступными для мозга, в то время как наблюдаемые состояния — являются доступными. Кроме того, тезис о том, что мозгом воспринимаются не сами объекты внешней среды непосредственно, а через сенсорику, то есть опосредованно или тезис о том, что между мозгом и окружающей средой есть своеобразная стена за которую мозг не может выйти и не может её преодолеть, однозначно указывают на кантианство фристоновской теории. Теперь, если сравнить это всё со следующей цитатой из Канта [128], то сомнения окончательно отпадают: «*... нам даны вещи как вне нас находящиеся предметы наших чувств, но о том, каковы они сами по себе, мы ничего не знаем, а знаем только их явления, то есть представления, которые они в нас производят, воздействуя на наши чувства. Следовательно, я, конечно, признаю, что вне нас существуют тела, то есть вещи, относительно которых нам совершенно неизвестно, каковы они сами по себе, но о которых мы знаем по представлениям, доставляемым нам их влиянием на нашу чувственность и получающим от нас название тел,— название, означающее, таким образом, только явление того неизвестного нам, но тем не менее действительного предмета...*». В случае же попытки некантианской интерпретации теории Фристона, например,

интерпретации в духе экологической психологии и прагматизма, предлагаемой Э. Кларком [16], [129],[130] и Дж. Брюнибергом и соавторами [131], кантианство из теории всё равно никуда не уходит, так как математический аппарат, вращающийся вокруг формулы Байеса и ненаблюдаемых и наблюдаемых состояний, в сущности, является перенесением кантианства на язык математики, где ненаблюдаемые состояния можно назвать «вещью в себе», а наблюдаемые состояния «вещью как явление». Исходя из этих соображений, теорию Фристона можно обозначить как «кантианство, изложенное на языке байесовской вероятности» или своеобразный «синтез кантианства и теоремы Байеса».

Проблема здесь заключается не в кантианстве как таковом, а в тех последствиях, которые оно влечет за собой. Дело в том, что кантианство дает отрицательный ответ на вопрос о познаваемости мира [74], [128], говоря о том, что объекты материального мира как «вещи в себе», недоступны для человеческого познания, для науки. Соответственно, научная теория, содержащая, пусть и неявно, в себе кантианство, автоматически поднимает вопросы о собственной научности и собственной пригодности для адекватного описания работы мозга, и сама же на него отвечает, по нашему мнению, в отрицательном смысле. Хотя стоит заметить, что некоторые философы-материалисты, в отличие от самого Канта, в вопросе познаваемости мира трактовали кантианство в положительном смысле [132]. Однако, даже и в этом случае теория оказывается под ударом многовековой критики кантианства, потому что та амфиболия, двусмысленность, которая, как неоднократно отмечалось философами [132], характерна для кантианства, по нашему мнению, свойственна и теории Фристона. Особенно это отчетливо это проявляется в вопросе о том, о чём же собственно делает гипотезы мозг: о ненаблюдаемых состояниях как таковых во всей полноте их характеристик или о каких-то атрибутах ненаблюдаемых состояний, где целостное познание этих состояний остается для мозга недоступным. Соответственно «тяжелый груз» критики кантианства ложится, как мы считаем, на любую теорию, допускающую в себе кантианство, какой, по нашему мнению, является теория Фристона.

2) *Марковость*. Это означает, что поведение случайных величин в теории не зависит от истории их динамики, а зависит только от прошлого значения. Требуется убедительное обоснование возможности моделирования работы мозга с помощью марковских процессов.

3) *Для функционирования «фристоновского мозга» в условиях многозадачности*, то есть в условиях, характерных для реального мозга, а не в тех искусственных и «игрушечных» условиях, которые характерны для моделей Фристона, *скорее потребуется некий универсальный метод оптимизации, хорошо применимый к любой модели мира.* Дело в том, что известно, что для одних оптимизационных задач хорошо работают одни методы, а другие — плохо [133], и универсального метода, скорее всего, просто не существует в принципе. Соответственно, встаёт вопрос о возможности дальнейшего развития развитии теории и, в целом, о её перспективности. Если потребуется такой универсальный метод оптимизации, то на наш взгляд, перспективы теории Фристона весьма сомнительны.

4) *Не указано, что же такое состояния в приложении к данной теории, и нет четкой связи состояний мозга с нейрональной активностью.* Фристон не дает определения такому широкому и неоднозначному понятию как понятию «состояния» в своей теории и пишет о том, что состояния связаны со скоростью нейрональной импульсации. К примеру, он пишет, что скорость нейрональной импульсации кодируется гипотезами о математическом ожидании ненаблюдаемых состояний [134]. Однако, в его работах нам не удалось найти ни одного упоминания о том, каким образом, например, можно прямо сопоставить модуль состояния конкретного уровня иерархии и, соответственно, гипотезы о математическом ожидании ненаблюдаемых состояний с ними связанных, с одной стороны, и активность (скорость импульсации) нейронов соответствующих этому модулю состояния, с другой стороны.

5) В теории Фристона *предполагается существование нейронных популяций, связанных с генерацией предсказаний и ошибок предсказаний*, однако *реальное наличие таких популяций не было продемонстрировано*, насколько нам известно, *ни в одном опыте.*

6) В теории Фристона *предполагается также циркуляция предсказаний и ошибок предсказаний по сетям нейронов, однако ничего подобного*, насколько нам известно, *не было показано на опыте.* А из пунктов 5 и 6, соответственно, следует *сомнительность гипотезы предиктивного кодирования.*

7) *Нет четкой привязки моделей Фристона к конкретным натурным экспериментам*, поэтому результаты моделирования не ясно к чему приложимы, а предсказания не ясно как сопоставлять с натурными экспериментальными данными. В целом, неясно, как

количественные и качественные предсказания теории могут быть сопоставлены с натурным экспериментом.

8) *При моделировании у Фристона используется модель какого-то отдельного процесса. Но для работоспособной порождающей модели, сравнимой с реальным мозгом, в идеале потребуется иметь модель ВСЕГО мира.*

9) *Сомнительны основы принципа свободной энергии*, то есть сомнительно представление о биологической системе как об эргодическом случайном динамическом аттракторе, а также в аргументации принципа свободной энергии происходит сомнительный перескок в рассуждениях из термодинамики в теорию информации. Об этом мы подробно писали в главе «Принцип свободной энергии».

10) Фристон пишет о том, что *в его теории большую роль имеет шум*, например, шум состояний [3], [5]. *Однако нам не известно работ, где бы было показано, что роль шума в мозговых процессах настолько велика*, как этого требует теория Фристона.

11) *Сомнительна гипотеза байесовского мозга.* Несмотря на то, что эксперименты показывают байес-оптимальность, которую можно обнаружить при анализе данных о восприятии [38], из этого вовсе не следует, что основу работы мозга составляют оценки по Байесу.

12) О форме неизвестных распределений. *Насколько правильно приближать неизвестные распределения именно гауссианом?* Убедительного обоснования для этого нет.

13) О линеаризации. *Насколько правильно отбрасывать нелинейные члены (степени, степени производных и т.д.) при представлении обобщенных координат?*

14) О приближении среднего поля. *Насколько правильно использовать приближение среднего поля?* На наш взгляд для этого нужны обоснования.

15) *Проблема априорной вероятности.* Фристон пишет, что иерархический подход позволяет решить проблему априорной вероятности, характерной для байесовской теории вероятностей [1]. По нашему мнению, даже если система иерархическая, то всё равно требуется задать исходную априорную вероятность на самом нижнем уровне иерархии. Соответственно, возникает вопрос, откуда берется такая априорная вероятность. Если же такая вероятность врожденная, то каким образом реализуется эта врожденность?

16) *Как непосредственно на натурном эксперименте количественно измерять или оценивать предсказание, ошибку предсказания, информационную свободную энергию и другие теоретико-информационные величины* теории Фристона?

Кроме того, отдельно стоит упомянуть ещё один серьёзный недостаток, присущий, как минимум, одной работе Фристона, но не являющийся недостатком теории как таковой. Это недостаточная обоснованность конструкций Фристона с точки зрения нейроанатомии. Чему примером является разобранное нами выше моделирование Фристоном спонтанного моргания, мигательного безусловного рефлекса и отставленного и следового условного рефлексов.

Таким образом, легко можно увидеть, что проблем, недостатков и неясностей в теории Фристона пока гораздо больше, чем достоинств. Эти проблемы мы можем разделить на 4 большие группы:

1) *философские* — кантианство,

2) *математические* — марковость, требование возможного универсального метода оптимизации для работоспособности теории Фристона в «реалистичных» ситуациях, линеаризация, форма неизвестных распределений, приближение среднего поля, природа исходной априорной вероятности,

3) *общие* — неопределенность понятия «состояния» и его связи с нейрональной активностью, то, что не было показано в эксперименте существование популяций, кодирующих ошибки предсказания и предсказания, а также циркуляции ошибок предсказания и предсказаний по нейрональным сетями, сомнительны основания теории, а также её неотъемлемые части: гипотезы байесовского мозга и предиктивного кодирования, проблема измерения или оценки в эксперименте ошибок предсказания, предсказаний, информационной свободной энергии и т.д.

4) *для моделирования* — отсутствует привязка моделей к конкретным экспериментальным данным.

9. Заключение

Теория Фристона, несмотря на свою оригинальность, универсальность и гибкость вызывает больше вопросов, чем дает ответов. Вызывают вопросы не только основания теории, допущения, в ней фигурирующие,

философские и математические проблемы, с ней связанные, но и её содержание, а также её предсказания, полезность и пригодность для моделирования реальных мозговых процессов.

Исходя из всего этого, по нашему мнению, следует признать, что теория Фристона пока ещё далека от окончательного своего вида и ещё до конца не сформулирована. По нашему мнению, она требует серьезной доработки и подробных пояснений в отношении упомянутых выше недостатков и вопросов, которые она вызывает.

Список литературы

[1] Friston K., "The free-energy principle: a unified brain theory?", *Nat. Rev. Neurosci.*, **11**:2 (2010), 127–38.

[2] Friston K.J., Trujillo-Barreto N., Daunizeau J., "DEM: a variational treatment of dynamic systems", *Neuroimage*, **41**:3 (2008), 849–885.

[3] Friston K., "Hierarchical Models in the Brain", *PLoS Comput. Biol.*, **4**:11 (2008), e1000211.

[4] Friston K.J., Daunizeau J., Kilner J., Kiebel S.J., "Action and behavior: a free-energy formulation", *Biol. Cybern.*, **102**:3 (2010), 227–260.

[5] Friston K., Kilner J., Harrison L., "A free energy principle for the brain", *J. Physiol. Paris.*, **100**:1–3 (2006), 70–87.

[6] Friston K.J., Daunizeau J., Kiebel S.J., "Reinforcement learning or active inference?", *PLoS One*, **4**:7 (2009), Article № e6421.

[7] Friston K., Stephan K., Li B., and Daunizeau J., "Generalised Filtering", *Mathematical Problems in Engineering*, **2010** (2010), 621670.

[8] Friston K., Mattout J., Trujillo-Barreto N., Ashburner J., Penny W., "Variational free energy and the Laplace approximation", *Neuroimage*, **34**:1 (2007), 220–234.

[9] Friston K., Herreros I., "Active Inference and Learning in the Cerebellum", *Neural Comput.*, **28**:9 (2016), 1812–1839.

[10] Kiebel S.J., Friston K.J., "Free energy and dendritic self-organization", *Front. Syst. Neurosci.*, **5** (2011), 80.

[11] Friston K., "A theory of cortical responses", *Philos. Trans. R Soc. Lond B Biol. Sci.*, **360**:1456 (2005), 815–836.

[12] Friston K., Schwartenbeck P., FitzGerald T., Moutoussis M., Behrens T., Dolan R.J., "The anatomy of choice: active inference and agency", *Front. Hum. Neurosci.*, **7** (2013), 598.

[13] von Helmholtz, H., *Handbuch der physiologischen Optik. 3.*, Leopold Voss, Leipzig, 1867, 874 c.

[14] Boring, E. G., *A history of experimental psychology*, Appleton-Century-Crofts, New York, 1950, 699 c.

[15] Helmholtz, H., *Concerning the perceptions in general. In Treatise on physiological optics (J. Southall, Trans., 3rd ed., Vol. III)*, Dover, New York, 1920-25, 734 c.

[16] Clark A., *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press, New York, 2015, 416 c.

[17] Hohwy J., *The Predictive Mind*, Oxford University Press, Oxford, 2014, 282 c.

[18] Swanson L. R., "The Predictive Processing Paradigm Has Roots in Kant", *Front. Syst. Neurosci.*, **10** (2016), 79

[19] Бэкон Ф., *Новый Органон*, Рипол классик, М., 2021, 364 с.

[20] Kassler, J. C., *The beginnings of the modern philosophy of music in England.*, Ashgate, Aldershot, 2004, 258 c.

[21] Gregory, R. L., "Perceptions as hypotheses.", *Phil. Trans. R Soc. Lond B.*, **290** (1980), 181–197

[22] Schneider, D. J., Hastorf, A. H., Ellsworth, P. C., *Person perception*, Addison-Wesley, Mass., 1979, 321 c.

[23] Lewicki, P., *Nonconscious social information processing*, Academic Press, New York, 1986, 237 c.

[24] Uleman J. S., Newman, L. S., Moskowitz, G. B., "People as flexible interpreters: Evidence and issues from spontaneous trait inference.", *Advances in Experimental Social Psychology,*, **28** (1996), 211–279

[25] Newman, L. S.; Uleman, J. S., *Spontaneous trait inference. In Uleman, J. S.; Bargh, J. A. (eds.). Unintended thought*, Guilford, New York, 1989, 155–188 c.

[26] Uleman, J. S.; Bargh, J. A., *Unintended thought*, Guilford, New York, 1989, 469 c.

[27] Hatfield G., *Perception as Unconscious Inference // Heyer D., Mausfeld R. Perception and the Physical World: Psychological and Philosophical Issues in Perception.*, John Wiley & Sons, Ltd,, 2002, 344 с.

[28] Boring, E. G., *Sensation and Perception in the History of Experimental Psychology*, Appleton-Century & Co, New York, 1942, 644 с.

[29] Dayan, P., Hinton, G. E., Neal, R., Zemel R. S., "The Helmholtz machine", *Neural Comput.*, **7**:5 (1995), 889–904

[30] Rao, R., Ballard, D., "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects", *Nat. Neurosci.*, **2**:1 (1999), 79–87

[31] Зюко А. Г., Кловский Д. Д., Назаров М. В., Финк Л., *Теория передачи сигналов: Учебник для вузов*, Связь, М., 1980, 288 с.

[32] Баринов С. М. и др., *Большой англо-русский политехнический словарь. В 2 т.: ок. 200000 терминов. Т. 1 : A-L = The comprehensive English- Russian scientific and technical dictionary*, Рус. яз., М., 1991, 701 с.

[33] Хокинс Д., Блейксли С., *Об интеллекте.*, Вильямс, М., 2007, 240 с.

[34] Carbajal G. V., Malmierca M. S., "The Neuronal Basis of Predictive Coding Along the Auditory Pathway: From the Subcortical Roots to Cortical Deviance Detection", *Trends Hear*, **22** (2018), № 2331216518784822

[35] Heilbron M., Chait M., "Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex?", *Neuroscience*, **389** (2018), 54–73

[36] Соколов, Е. Н., *Восприятие и условный рефлекс: Новый взгляд.*, УМК «Психология»; Московский психолого-социальный институт, М., 2003, 287 с.

[37] Barron H. C., Auksztulewicz R., Friston K., "Prediction and memory: A predictive coding account", *Prog. Neurobiol.*, **192** (2020), № 101821

[38] Knill D.C., Pouget A., "The Bayesian brain: the role of uncertainty in neural coding and computation", *Trends Neurosci.*, **27**:12 (2004), 712–719

[39] Friston K., Schwartenbeck P., FitzGerald T., Moutoussis M., Behrens T., Dolan R. J., "The anatomy of choice: dopamine and decision-making", *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **369**:1655 (2014), № 20130481

[40] Meera A. A. and Wisse M., “Free energy principle based state and input observer design for linear systems with colored noise”, *2020 American Control Conference (ACC). IEEE*, 2020, 5052–5058.

[41] Meera A. A. and Wisse M., “Free Energy Principle for the Noise Smoothness Estimation of Linear Systems with Colored Noise”, *ArXiv:2204.01796*, 2022

[42] Karl Friston, https://scholar.google.co.uk/ citations?user=q_ 4u0aoAAAAJ&hl=en (дата обращения: 18.11.2022).

[43] Friston K. J., Stephan K. E., “Free-energy and the brain”, *Synthese*, **159**:3 (2007), 417–458.

[44] Friston K., Ao P., “Free energy, value, and attractors”, *Comput. Math. Methods Med.*, **2012** (2012), № 937860.

[45] Aguilera M., Millidge B., Tschantz A., Buckley C. L., “How particular is the physics of the free energy principle?”, *Phys. Life Rev.*, **40** (2022), 24–50.

[46] Mazzaglia P., Verbelen T., Çatal O., Dhoedt B., “The Free Energy Principle for Perception and Action: A Deep Learning Perspective”, *Entropy (Basel)*, **24**:2 (2022), 301.

[47] Raja V., Valluri D., Baggs E., Chemero A., Anderson M. L., “The Markov blanket trick: On the scope of the free energy principle and active inference”, *Phys. Life Rev.*, **39** (2021), 49–72.

[48] Boyadzhieva A., Kayhan E., “Keeping the Breath in Mind: Respiration, Neural Oscillations, and the Free Energy Principle”, *Front Neurosci.*, **15** (2021), № 647579.

[49] Friston K., “A Free Energy Principle for Biological Systems”, *Entropy (Basel)*, **14**:11 (2012), 2100–2121.

[50] Бернштейн Н. А., *Пути и задачи физиологии активности. //Бернштейн Н. А. Физиология движений и активность.*, М., 1990, 438 с.

[51] Шрёдингер Э., *Что такое жизнь? Физический аспект живой клетки*, Ижевск: НИЦ «Регулярная и хаотическая динамика, М., 2002, 92 с.

[52] Больцман Л., *Избранные труды. Молекулярно-кинетическая теория газов. Термодинамика. Статистическая механика. Теория излучения. Общие вопросы физики.*, Наука, М., 1984, 590 с.

[53] Brownlee J., *Probability for Machine Learning: Discover How To Harness Uncertainty With Python*, Machine Learning Mastery., 2020, 319 с.

[54] Верещагин Н. К., Щепин Е. В., *Информация, кодирование и предсказание.*, ФМОП, В31 МЦНМО, М., 2012, 236 с.

[55] Murphy K. P., *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012, 1104 с.

[56] Ландау Л. Д., Лифшиц Е. М., *Статистическая физика. Часть 1: Учебное пособие для вузов.*, Физматлит, М., 2010, 616 с.

[57] Hinton G. E., van Camp D., "Keeping the neural networks simple by minimizing the description length of the weights", *COLT*, 1993, 5–13

[58] Hinton, G. E. and Zemel, R. S., *Autoencoders, Minimum Description Length, and Helmholtz Free Energy. Cowan J. D., Tesauro G., Alspector J. Advances in Neural Information Processing Systems 6.*, Morgan Kaufmann, San Mateo, CA, 1994, 3–10 с.

[59] Zemel, R. S., Hinton, G. E., "Learning Population Codes by Minimizing Description Length", *Neural Computation*, **7**:3 (1995), 549–564

[60] Hinton G. E., Dayan P., Frey B. J., Neal R. M., "The "wake-sleep"algorithm for unsupervised neural networks", *Science*, **268**:5214 (1995), 1158–1161

[61] Сена Л. А., *Единицы физических величин и их размерности*, Наука, М., 1988, 432 с.

[62] Colombo, M., Palacios, P., "Non-equilibrium thermodynamics and the free energy principle in biology", *Biol Philos*, **36** (2021), 41

[63] Biehl M., Pollock F., Kanai R., "A Technical Critique of Some Parts of the Free Energy Principle", *Entropy*, **23**:3 (2021), 293

[64] Friston K. J., Da Costa L., Parr T., "Some Interesting Observations on the Free Energy Principle", *Entropy (Basel)*, **23**:8 (2021), 1076

[65] Friston K., Heins C., Ueltzhöffer K., Da Costa L., Parr T., "Stochastic Chaos and Markov Blankets", *Entropy (Basel)*, **23**:9 (2021), 1220

[66] Ландау Л. Д., Лифшиц Е. М., *Механика. 5-е изд., стереотип.*, Физматлит, М., 2012, 224 с.

[67] Ландау Л. Д., Лифшиц Е. М., *Теория поля. Издание 8-е, стереотип.*, Физматлит, М., 2012, 536 с.

[68] Friston K., “A free energy principle for a particular physics”, *ArXiv:190610184*, 2019

[69] Li S., Zhuang C., Hao M., He X., Marquez J. C., Niu C. M., Lan N., “Coordinated alpha and gamma control of muscles and spindles in movement and posture”, *Front. Comput. Neurosci.*, **9** (2015), 122

[70] Lopez-Poveda E. A., “Olivocochlear Efferents in Animals and Humans: From Anatomy to Clinical Relevance”, *Front Neurol.*, **9** (2018), 197

[71] Gastinger M. J., Tian N., Horvath T., Marshak D. W., “Retinopetal Axons in Mammals: Emphasis on Histamine and Serotonin”, *Curr Eye Res.*, **31**:7–8 (2006), 1655–1667

[72] Adams, R.A., Shipp, S., Friston, K.J., “Predictions not commands: active inference in the motor system”, *Brain Struct. Funct.*, **218**:3 (2013), 611–643

[73] Платон, *Собрание сочинений. В 4 т. / Под общ. ред. А. Ф. Лосева, В. Ф. Асмуса, А. А. Тахо-Годи. (Серия «Философское наследие»). Т. 2.*, Мысль, М., 1993, 528 с.

[74] Кант И., *Критика чистого разума / Пер. с нем. Н. Лосского сверен и отредактирован Ц. Г. Арзаканяном и М. И. Иткиным; Примеч. Ц. Г. Арзаканяна.*, Эксмо, М., 2007, 736 с.

[75] Фейнман Р., Лейтон Р., Сендс М., *Современная наука о природе. Законы механики. Пространство. Время. Движение. Т. 1.*, АСТ, М., 2019, 478 с.

[76] Звягин Л.С., “Перспективы применения моделей с латентными переменными и вариационный байес для задач оптимизации”, *Международная конференция по мягким вычислениям и измерениям*, **1** (2019), 25–29

[77] Савин А.В., “Байесовский подход в современном анализе: алгоритмы и синтез”, *Международная конференция по мягким вычислениям и измерениям*, **2** (2018), 635–638

[78] Баранов И.Д., “Байесовская стратегия оценки достоверности и метод байесовских сетей”, *Международная конференция по мягким вычислениям и измерениям*, **2** (2018), 659–662

[79] Marinai S., Fujisawa H., *Machine Learning in Document Analysis and Recognition*, Springer-Verlag Berlin Heidelberg, Berlin, 2008, 434 с.

[80] Srinivasan H., Srihari S. N., Beal M. J., “Machine Learning for Signature Verification”, *// Kalra P., Peleg S. Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006, Madurai, India, December 13-16, 2006*, 2007, 761–775

[81] Флах, П., *Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных*, ДМК ПРЕСС, М., 2015, 399 с.

[82] Friston K., Breakspear M., Deco G., “Perception and self-organized instability”, *Front. Comput. Neurosci.*, **6** (2012), 44

[83] Friston K., Kiebel S., “Predictive coding under the free-energy principle”, *Philos Trans R Soc Lond B Biol Sci.*, **364**:1521 (2009), 1211–1221

[84] Parr T., Rees G., Friston K. J., “Computational Neuropsychology and Bayesian Inference”, *Front. Hum. Neurosci.*, **12** (2018), 61

[85] Friston K., Parr T., Zeidman P., “Bayesian model reduction”, *ArXiv:1805.07092.*, 2018

[86] Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques, 3rd ed.*, Elsevier, New York, 2012, 744 с.

[87] McCullagh P., “What is a statistical model?”, *Ann. Statist.*, **30**:5 (2002), 1225–1310

[88] Кадис Р. Л., “О терминах и понятиях "точность"и "правильность"(результатов) химического анализа”, *Журнал аналитической химии*, **62**:6 (2007), 566–574

[89] Statistical parametric mapping. URL: https://www.fil.ion.ucl.ac.uk/spm/ (дата обращения: 18.11.2022).

[90] Федорюк М. В., *Метод перевала*, URSS, М., 2015, 368 с.

[91] Bishop C. M., *Pattern Recognition and Machine Learning*, Springer New York, New York, 2006, 738 с.

[92] Felleman D. J., Van Essen D. C., “Distributed hierarchical processing in the primate cerebral cortex”, *Cereb Cortex*, **1**:1 (1991), 1–47

[93] Bastos A. M., Usrey W. M., Adams R. A., Mangun G. R., Fries P., Friston K. J., “Canonical microcircuits for predictive coding”, *Neuron*, **76**:4 (2012), 695–711

[94] Ozaki T., "A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach", *Statistica Sinica*, **2** (1992), 113–135

[95] Balaban C.D., "Olivo-vestibular and cerebello-vestibular connections in albino rabbits", *Neuroscience*, **12**:1 (1984), 129–149

[96] Highstein S.M., Holstein G.R., "The anatomy of the vestibular nuclei", *Prog Brain Res.*, **151** (2006), 157-203

[97] Basso M. A., Evinger C., "An Explanation for Reflex Blink Hyperexcitability in Parkinson's Disease. II. Nucleus Raphe Magnus", *J. Neurosci.*, **16**:22 (1996), 7318–7330

[98] Peterson D. A., Sejnowski T. J., "A Dynamic Circuit Hypothesis for the Pathogenesis of Blepharospasm", *Front Comput Neurosci.*, **11** (2017), 11

[99] Baumel Y., Jacobson G.A., Cohen D., "Implications of functional anatomy on information processing in the deep cerebellar nuclei", *Front. Cell. Neurosci.*, **3** (2009), 14

[100] Friston K. J., Shiner T., FitzGerald T., Galea J. M., Adams R., Brown H., Dolan R. J., Moran R., Stephan K. E., Bestmann S., "Dopamine, Affordance and Active Inference", *PLoS Comput Biol.*, **8**:1 (2012), № e1002327

[101] Parr T., Friston K. J., "The Anatomy of Inference: Generative Models and Brain Structure", *Front Comput Neurosci.*, **12** (2018), 90

[102] Carhart-Harris R. L., Friston K. J., "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics", *Pharmacol. Rev.*, **71**:3 (2019), 316–344

[103] Aramideh M., Ongerboer de Visser B. W., Koelman J. H., Majoie C. B., Holstege G., "The late blink reflex response abnormality due to lesion of the lateral tegmental field", *Brain*, **120**:9 (1997), 1685–1692

[104] Медведева Л.А., Сыровегин А.В., Авакян Г.Н., Гнездилов А.В., Загорулько О.И., "Методология исследования мигательного рефлекса и его нормативные параметры", *Журнал неврологии и психиатрии*, **1** (2011), 62–67

[105] Boele H.-J., Koekkoek S. K. E., De Zeeuw C. I., "Cerebellar and extracerebellar involvement in mouse eyeblink conditioning: the ACDC model", *Front. Cell Neurosci.*, **3** (2010), 19

[106] Freeman J. H., Steinmetz A. B., "Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning", *Learn Mem.*, **18**:10 (2011), 666–677

[107] Yang Y., Lei C., Feng H., Sui J.-A., "The neural circuitry and molecular mechanisms underlying delay and trace eyeblink conditioning in mice", *Behav. Brain Res.*, **278** (2015), 307–314

[108] Jirenhed D.-A., Bengtsson F., Hesslow G., "Acquisition, extinction, and reacquisition of a cerebellar cortical memory trace", *J. Neurosci.*, **27**:10 (2007), 2493–2502

[109] Ito M., Kano M., "Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex", *Neurosci. Lett.*, **33**:3 (1982), 253–258

[110] Ito M., *The Cerebellum and Neural Control*, Raven, New York, 1984, 580 c.

[111] Mandwal A., Orlandi J. G., Simon C., Davidsen J., "A biochemical mechanism for time-encoding memory formation within individual synapses of Purkinje cells", *PLoS One*, **16**:5 (2021), № e0251172

[112] Delgado-Garcia J. M., Gruart A., "The role of interpositus nucleus in eyelid conditioned responses", *Cerebellum*, **1**:4 (2002), 289–308

[113] Delgadoz-Garcia J. M., Gruart A., "Firing activities of identified posterior interpositus nucleus neurons during associative learning in behaving cats", *Brain Res. Rev.*, **49**:2 (2005), 367–376

[114] Delgado-Garcia J. M., Gruart A., "Building new motor responses: eyelid conditioning revisited", *Trends Neurosci.*, **29**:6 (2006), 330–338

[115] Aksenov D., Serdyukova N., Irwin K., Bracha V., "GABA neurotransmission in the cerebellar interposed nuclei: involvement in classically conditioned eyeblinks and neuronal activity", *J. Neurophysiol.*, **91**:2 (2004), 719–727

[116] Aksenov D. P., Serdyukova N. A., Bloedel J. R., Bracha V., "Glutamate neurotransmission in the cerebellar interposed nuclei: involvement in classically conditioned eyeblinks and neuronal activity", *J. Neurophysiol.*, **93**:1 (2005), 44–52

[117] Blankenship M. R., Huckfeldt R., Steinmetz J. J., Steinmetz J. E., "The effects of amygdala lesions on hippocampal activity and classical eyeblink conditioning in rats", *Brain Res.*, **1035**:2 (2005), 120–130

[118] Weisz D. J., Harden D. G., Xiang Z., “Effects of amygdala lesions on reflex facilitation and conditioned response acquisition during nictitating membrane response conditioning in rabbit”, *Behav. Neurosci.*, **106**:2 (1992), 262–273

[119] Lee T., Kim J., “Differential effects of cerebellar, amygdalar, and hippocampal lesions on classical eyeblink conditioning in rats”, *J. Neurosci.*, **24**:13 (2004), 3242–3250

[120] Гришаев А.В., Сазонов В.Ф., “Роль активных и пассивных свойств дендритного дерева нейрона в многоплановой интеграции постсинаптических потенциалов”, *Усп. физиол. наук*, **51**:3 (2020), 87–104

[121] Сазонов В.Ф., Сазонов И.В., Гришаев А.В., “Понятие режимности в работе нейронов как функциональная альтернатива структурной пластичности в компьютерном моделировании межнейронных взаимодействий”, *Нейрокомпьютеры: разработка, применение*, **22**:5 (2020), 43–53

[122] Гришаев А.В., Сазонов В.Ф., “Дендритные механизмы регуляции активности и пластичности пирамидных нейронов неокортекса”, *Усп. физиол. наук.*, **49**:3 (2018), 104–118

[123] Friston K., Kiebel S., “Cortical circuits for perceptual inference”, *Neural Netw.*, **22**:8 (2009), 1093–1104

[124] Clark, A., “Whatever next? Predictive brains, situated agents, and the future of cognitive science”, *Behav. Brain Sci.*, **36**:3 (2013), 181–204

[125] Gładziejewski, P., “Predictive coding and representationalism”, *Synthese*, **193** (2016), 559–582

[126] Hohwy, J., *The Predictive Mind*, Oxford University Press, New York, 2013, 288 с.

[127] Hohwy, J., “The self-evidencing brain”, *Noûs*, **50**:2 (2014), 259–285

[128] Кант И., *Пролегомены ко всякой будущей метафизике, могущей появиться как наука.// Кант И. Сочинения в шести томах. Т. 4, Ч. 1.*, М., 1965, 544 с.

[129] Clark, A., *Predicting peace: the end of the representation wars. In: Metzinger T., Windt J. M. Frankfurt am Main: MIND Group*, MIND Group, Frankfurt am Main, 2015, 1–7 с.

[130] Clark, A., “Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil”, *Noûs*, **51**:4 (2016), 727–753

[131] Bruineberg, J., Kiverstein, J., Rietveld, E., "The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective", *Synthese*, **195**:6 (2018), 2417–2444

[132] Ленин В. И., *Полное собрание сочинений: в 55 т. / В. И. Ленин; Ин-т марксизма-ленинизма при ЦК КПСС — 5-е изд. Т. 18. Материализм и эмпириокритицизм.*, Гос. изд-во полит. лит., М., 1968, 7—384 с.

[133] Yang X.-S., *Optimization Techniques and Applications with Examples*, John Wiley & Sons, Inc. Hoboken, New Jersey, 2018, 384 c.

[134] Friston K., FitzGerald T., Rigoli F., Schwartenbeck P., Pezzulo G., "Active Inference: A Process Theory", *Neural Comput.*, **29**:1 (2017), 1–49

The Karl Friston's neurobiological theory: a critical review
Grishaev A.V., Sazonov V.F.

In this article, for the first time in the Russian-language literature, the neurobiological version of Karl Friston's theory is presented. For the first time, this theory is presented holistically, in detail and logically built within the framework of one article and, as far as possible, adapted for understanding by neuroscientists and neurophysiologists. How Friston applies this theory in practice is discussed. A critical analysis of internal problems and contradictions in Friston's theory is carried out.

Keywords: free energy principle, Bayesian brain hypothesis, predictive coding, prediction error, observed states, unobserved states.

References

[1] Friston K., "The free-energy principle: a unified brain theory?", *Nat. Rev. Neurosci.*, **11**:2 (2010), 127–38

[2] Friston K.J., Trujillo-Barreto N., Daunizeau J., "DEM: a variational treatment of dynamic systems", *Neuroimage*, **41**:3 (2008), 849–885

[3] Friston K., "Hierarchical Models in the Brain", *PLoS Comput. Biol.*, **4**:11 (2008), e1000211

[4] Friston K.J., Daunizeau J., Kilner J., Kiebel S.J., "Action and behavior: a free-energy formulation", *Biol. Cybern.*, **102**:3 (2010), 227–260

[5] Friston K., Kilner J., Harrison L., "A free energy principle for the brain", *J. Physiol. Paris.*, **100**:1–3 (2006), 70–87

[6] Friston K.J., Daunizeau J., Kiebel S.J., "Reinforcement learning or active inference?", *PLoS One*, **4**:7 (2009), Article № e6421

[7] Friston K., Stephan K., Li B., and Daunizeau J., "Generalised Filtering", *Mathematical Problems in Engineering*, **2010** (2010), 621670

[8] Friston K., Mattout J., Trujillo-Barreto N., Ashburner J., Penny W., "Variational free energy and the Laplace approximation", *Neuroimage*, **34**:1 (2007), 220–234

[9] Friston K., Herreros I., "Active Inference and Learning in the Cerebellum", *Neural Comput.*, **28**:9 (2016), 1812–1839

[10] Kiebel S.J., Friston K.J., "Free energy and dendritic self-organization", *Front. Syst. Neurosci.*, **5** (2011), 80

[11] Friston K., "A theory of cortical responses", *Philos. Trans. R Soc. Lond B Biol. Sci.*, **360**:1456 (2005), 815–836

[12] Friston K., Schwartenbeck P., FitzGerald T., Moutoussis M., Behrens T., Dolan R.J., "The anatomy of choice: active inference and agency", *Front. Hum. Neurosci.*, **7** (2013), 598

[13] von Helmholtz, H., *Handbuch der physiologischen Optik. 3.*, Leopold Voss, Leipzig, 1867, 874 c.

[14] Boring, E. G., *A history of experimental psychology*, Appleton-Century-Crofts, New York, 1950, 699 c.

[15] Helmholtz, H., *Concerning the perceptions in general. In Treatise on physiological optics (J. Southall, Trans., 3rd ed., Vol. III)*, Dover, New York, 1920-25, 734 c.

[16] Clark A., *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press, New York, 2015, 416 c.

[17] Hohwy J., *The Predictive Mind*, Oxford University Press, Oxford, 2014, 282 c.

[18] Swanson L. R., "The Predictive Processing Paradigm Has Roots in Kant", *Front. Syst. Neurosci.*, **10** (2016), 79

[19] Bacon F., *New Organon*, Ripol classic, M., 2021 (In Russian), 364 c.

[20] Kassler, J. C., *The beginnings of the modern philosophy of music in England.*, Ashgate, Aldershot, 2004, 258 c.

[21] Gregory, R. L., "Perceptions as hypotheses.", *Phil. Trans. R Soc. Lond B.*, **290** (1980), 181–197

[22] Schneider, D. J., Hastorf, A. H., Ellsworth, P. C., *Person perception*, Addison-Wesley, Mass., 1979, 321 c.

[23] Lewicki, P., *Nonconscious social information processing*, Academic Press, New York, 1986, 237 c.

[24] Uleman J. S., Newman, L. S., Moskowitz, G. B., "People as flexible interpreters: Evidence and issues from spontaneous trait inference.", *Advances in Experimental Social Psychology,*, **28** (1996), 211–279

[25] Newman, L. S.; Uleman, J. S., *Spontaneous trait inference. In Uleman, J. S.; Bargh, J. A. (eds.). Unintended thought*, Guilford, New York, 1989, 155–188 c.

[26] Uleman, J. S.; Bargh, J. A., *Unintended thought*, Guilford, New York, 1989, 469 c.

[27] Hatfield G., *Perception as Unconscious Inference // Heyer D., Mausfeld R. Perception and the Physical World: Psychological and Philosophical Issues in Perception.*, John Wiley & Sons, Ltd,, 2002, 344 c.

[28] Boring, E. G., *Sensation and Perception in the History of Experimental Psychology*, Appleton-Century & Co, New York, 1942, 644 c.

[29] Dayan, P., Hinton, G. E., Neal, R., Zemel R. S., "The Helmholtz machine", *Neural Comput.*, **7**:5 (1995), 889–904

[30] Rao, R., Ballard, D., "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects", *Nat. Neurosci.*, **2**:1 (1999), 79–87

[31] Zyuko A. G., Klovskiy D. D., Nazarov M. V., Fink L., *Signaling Theory: A Textbook for Universities*, Communication, M., 1980 (In Russian), 288 c.

[32] Barinov S. M. et. al., *The comprehensive English- Russian scientific and technical dictionary. In two vol.: approx. 200000 entries. V. 1 : A-L*, Russ. yaz. publish., M., 1991 (In Russian), 701 c.

[33] Hawkins J., Blakeslee S., *On Intelligence.*, Times Books, New York, 2004, 272 c.

[34] Carbajal G. V., Malmierca M. S., “The Neuronal Basis of Predictive Coding Along the Auditory Pathway: From the Subcortical Roots to Cortical Deviance Detection”, *Trends Hear*, **22** (2018), № 2331216518784822

[35] Heilbron M., Chait M., “Great Expectations: Is there Evidence for Predictive Coding in Auditory Cortex?”, *Neuroscience*, **389** (2018), 54–73

[36] Sokolov, E. N., *Perception and conditioned reflex: A new look.*, UMK "Psychology"; Moscow Psychological and Social Institute, M., 2003 (In Russian), 287 c.

[37] Barron H. C., Auksztulewicz R., Friston K., “Prediction and memory: A predictive coding account”, *Prog. Neurobiol.*, **192** (2020), № 101821

[38] Knill D.C., Pouget A., “The Bayesian brain: the role of uncertainty in neural coding and computation”, *Trends Neurosci.*, **27**:12 (2004), 712–719

[39] Friston K., Schwartenbeck P., FitzGerald T., Moutoussis M., Behrens T., Dolan R. J., “The anatomy of choice: dopamine and decision-making”, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **369**:1655 (2014), № 20130481

[40] Meera A. A. and Wisse M., “Free energy principle based state and input observer design for linear systems with colored noise”, *2020 American Control Conference (ACC). IEEE*, 2020, 5052–5058.

[41] Meera A. A. and Wisse M., “Free Energy Principle for the Noise Smoothness Estimation of Linear Systems with Colored Noise”, *ArXiv:2204.01796*, 2022

[42] Karl Friston, https://scholar.google.co.uk/ citations?user=q_ 4u0aoAAAAJ&hl=en (date of the application: 18.11.2022).

[43] Friston K. J., Stephan K. E., “Free-energy and the brain”, *Synthese*, **159**:3 (2007), 417–458.

[44] Friston K., Ao P., “Free energy, value, and attractors”, *Comput. Math. Methods Med.*, **2012** (2012), № 937860.

[45] Aguilera M., Millidge B., Tschantz A., Buckley C. L., “How particular is the physics of the free energy principle?”, *Phys. Life Rev.*, **40** (2022), 24–50.

[46] Mazzaglia P., Verbelen T., Çatal O., Dhoedt B., “The Free Energy Principle for Perception and Action: A Deep Learning Perspective”, *Entropy (Basel)*, **24**:2 (2022), 301.

[47] Raja V., Valluri D., Baggs E., Chemero A., Anderson M. L., “The Markov blanket trick: On the scope of the free energy principle and active inference”, *Phys. Life Rev.*, **39** (2021), 49–72.

[48] Boyadzhieva A., Kayhan E., “Keeping the Breath in Mind: Respiration, Neural Oscillations, and the Free Energy Principle”, *Front Neurosci.*, **15** (2021), № 647579.

[49] Friston K., “A Free Energy Principle for Biological Systems”, *Entropy (Basel)*, **14**:11 (2012), 2100–2121.

[50] Bernstein N. A., *Ways and tasks of activity physiology. // Bernstein N. A. Physiology of movements and activity.*, M., 1990, 438 c.

[51] Schrodinger E., *What Is Life? The Physical Aspect of the Living Cell*, Izhevsk: Research Center "Regular and Chaotic Dynamics, M., 2002 (In Russian), 92 c.

[52] Boltzmann L., *Selected works. Molecular-kinetic theory of gases. Thermodynamics. Statistical mechanics. Theory of radiation. General questions of physics.*, Nauka, M., 1984 (In Russian), 590 c.

[53] Brownlee J., *Probability for Machine Learning: Discover How To Harness Uncertainty With Python*, Machine Learning Mastery., 2020, 319 c.

[54] Vereshchagin N. K., Shchepin E. V., *Information, coding and prediction.*, FMOP, V31 MCNMO, M., 2012 (In Russian), 236 c.

[55] Murphy K. P., *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012, 1104 c.

[56] Landau L. D., Lifshits E. M., *Statistical physics. Part 1: Textbook for universities.*, Fizmatlit, M., 2010 (In Russian), 616 c.

[57] Hinton G. E., van Camp D., “Keeping the neural networks simple by minimizing the description length of the weights”, *COLT*, 1993, 5–13

[58] Hinton, G. E. and Zemel, R. S., *Autoencoders, Minimum Description Length, and Helmholtz Free Energy. Cowan J. D., Tesauro G., Alspector J. Advances in Neural Information Processing Systems 6.*, Morgan Kaufmann, San Mateo, CA, 1994, 3–10 c.

[59] Zemel, R. S., Hinton, G. E., “Learning Population Codes by Minimizing Description Length”, *Neural Computation*, **7**:3 (1995), 549–564

[60] Hinton G. E., Dayan P., Frey B. J., Neal R. M., “The "wake-sleep"algorithm for unsupervised neural networks”, *Science*, **268**:5214 (1995), 1158–1161

[61] Sena L.A., *Units of physical quantities and their dimensions*, Nauka, M., 1988 (In Russian), 432 c.

[62] Colombo, M., Palacios, P., “Non-equilibrium thermodynamics and the free energy principle in biology”, *Biol Philos*, **36** (2021), 41

[63] Biehl M., Pollock F., Kanai R., “A Technical Critique of Some Parts of the Free Energy Principle”, *Entropy*, **23**:3 (2021), 293

[64] Friston K. J., Da Costa L., Parr T., “Some Interesting Observations on the Free Energy Principle”, *Entropy (Basel)*, **23**:8 (2021), 1076

[65] Friston K., Heins C., Ueltzhöffer K., Da Costa L., Parr T., “Stochastic Chaos and Markov Blankets”, *Entropy (Basel)*, **23**:9 (2021), 1220

[66] Landau L. D., Lifshits E. M., *Mechanics. 5th ed.*, Fizmatlit, M., 2012 (In Russian), 224 c.

[67] Landau L. D., Lifshits E. M., *Field theory. 8-th ed.*, Fizmatlit, M., 2012 (In Russian), 536 c.

[68] Friston K., “A free energy principle for a particular physics”, *ArXiv:190610184*, 2019

[69] Li S., Zhuang C., Hao M., He X., Marquez J. C., Niu C. M., Lan N., “Coordinated alpha and gamma control of muscles and spindles in movement and posture”, *Front. Comput. Neurosci.*, **9** (2015), 122

[70] Lopez-Poveda E. A., “Olivocochlear Efferents in Animals and Humans: From Anatomy to Clinical Relevance”, *Front Neurol.*, **9** (2018), 197

[71] Gastinger M. J., Tian N., Horvath T., Marshak D. W., “Retinopetal Axons in Mammals: Emphasis on Histamine and Serotonin”, *Curr Eye Res.*, **31**:7–8 (2006), 1655–1667

[72] Adams, R.A., Shipp, S., Friston, K.J., “Predictions not commands: active inference in the motor system”, *Brain Struct. Funct.*, **218**:3 (2013), 611–643

[73] Plato, *Collected works. In 4 volumes / Under the total. ed. A. F. Losev, V. F. Asmus, A. A. Takho-Godi. (Series "Philosophical heritage"). T. 2.*, Mysl, M., 1993 (In Russian), 528 c.

[74] Kant I., *Critique of Pure Reason / Trans. with ger. by N. Lossky verified and edited by Ts. G. Arzakanyan and M. I. Itkin; Note. Ts. G. Arzakanyan.*, Eksmo, M., 2007 (In Russian), 736 c.

[75] Feynman R., Layton R., Sands M., *Modern science of nature. The laws of mechanics. Space. Time. Movement. T. 1.*, AST, M., 2019 (In Russian), 478 c.

[76] Zvyagin L.S., "Prospects of application of models with latent variables and variational bayes for optimization problems", *International Conference on Soft Computing and Measurements*, **1** (2019), 25–29 (In Russian)

[77] Savin A.V., "Bayesian approach in modern analysis: algorithms and synthesis", *International Conference on Soft Computing and Measurements*, **2** (2018), 635–638 (In Russian)

[78] Baranov I.D., "Bayesian Reliability Estimation Strategy and Bayesian Network Method", *International Conference on Soft Computing and Measurements*, **2** (2018), 659–662 (In Russian)

[79] Marinai S., Fujisawa H., *Machine Learning in Document Analysis and Recognition*, Springer-Verlag Berlin Heidelberg, Berlin, 2008, 434 c.

[80] Srinivasan H., Srihari S. N., Beal M. J., "Machine Learning for Signature Verification", *// Kalra P., Peleg S. Computer Vision, Graphics and Image Processing: 5th Indian Conference, ICVGIP 2006, Madurai, India, December 13-16, 2006*, 2007, 761–775

[81] Flakh, П., *Machine learning. The science and art of building algorithms that extract knowledge from data*, DMK PRESS, M., 2015 (In Russian), 399 c.

[82] Friston K., Breakspear M., Deco G., "Perception and self-organized instability", *Front. Comput. Neurosci.*, **6** (2012), 44

[83] Friston K., Kiebel S., "Predictive coding under the free-energy principle", *Philos Trans R Soc Lond B Biol Sci.*, **364**:1521 (2009), 1211–1221

[84] Parr T., Rees G., Friston K. J., "Computational Neuropsychology and Bayesian Inference", *Front. Hum. Neurosci.*, **12** (2018), 61

[85] Friston K., Parr T., Zeidman P., “Bayesian model reduction”, *ArXiv:1805.07092.*, 2018

[86] Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques, 3rd ed.*, Elsevier, New York, 2012, 744 c.

[87] McCullagh P., “What is a statistical model?”, *Ann. Statist.*, **30**:5 (2002), 1225–1310

[88] Kadis R.L., “The terms "tochnost’"and "pravil’nost’"as applied to the results of chemical analysis”, *Journal of Analytical Chemistry.*, **62**:6 (2007), 566–574 (In Russian)

[89] Statistical parametric mapping. URL: https://www.fil.ion.ucl.ac.uk/spm/ (date of the application: 18.11.2022).

[90] Fedoryuk M.V., *Pass method*, URSS, M., 2015 (In Russian), 368 c.

[91] Bishop C. M., *Pattern Recognition and Machine Learning*, Springer New York, New York, 2006, 738 c.

[92] Felleman D. J., Van Essen D. C., “Distributed hierarchical processing in the primate cerebral cortex”, *Cereb Cortex*, **1**:1 (1991), 1–47

[93] Bastos A. M., Usrey W. M., Adams R. A., Mangun G. R., Fries P., Friston K. J., “Canonical microcircuits for predictive coding”, *Neuron*, **76**:4 (2012), 695–711

[94] Ozaki T., “A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach”, *Statistica Sinica*, **2** (1992), 113–135

[95] Balaban C.D., “Olivo-vestibular and cerebello-vestibular connections in albino rabbits”, *Neuroscience*, **12**:1 (1984), 129–149

[96] Highstein S.M., Holstein G.R., “The anatomy of the vestibular nuclei”, *Prog Brain Res.*, **151** (2006), 157-203

[97] Basso M. A., Evinger C., “An Explanation for Reflex Blink Hyperexcitability in Parkinson’s Disease. II. Nucleus Raphe Magnus”, *J. Neurosci.*, **16**:22 (1996), 7318–7330

[98] Peterson D. A., Sejnowski T. J., “A Dynamic Circuit Hypothesis for the Pathogenesis of Blepharospasm”, *Front Comput Neurosci.*, **11** (2017), 11

[99] Baumel Y., Jacobson G.A., Cohen D., "Implications of functional anatomy on information processing in the deep cerebellar nuclei", *Front. Cell. Neurosci.*, **3** (2009), 14

[100] Friston K. J., Shiner T., FitzGerald T., Galea J. M., Adams R., Brown H., Dolan R. J., Moran R., Stephan K. E., Bestmann S., "Dopamine, Affordance and Active Inference", *PLoS Comput Biol.*, **8**:1 (2012), № e1002327

[101] Parr T., Friston K. J., "The Anatomy of Inference: Generative Models and Brain Structure", *Front Comput Neurosci.*, **12** (2018), 90

[102] Carhart-Harris R. L., Friston K. J., "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics", *Pharmacol. Rev.*, **71**:3 (2019), 316–344

[103] Aramideh M., Ongerboer de Visser B. W., Koelman J. H., Majoie C. B., Holstege G., "The late blink reflex response abnormality due to lesion of the lateral tegmental field", *Brain*, **120**:9 (1997), 1685–1692

[104] Medvedeva L.A., Syrovegin A.V., Avakian G.N., Gnezdilov A.V., Zagorul'ko O.I., "The methodology on the study of blink reflex and its normative parameters", *S.S. Korsakov journal of neurology and psychiatry*, **1** (2011), 62–67 (In Russian)

[105] Boele H.-J., Koekkoek S. K. E., De Zeeuw C. I., "Cerebellar and extracerebellar involvement in mouse eyeblink conditioning: the ACDC model", *Front. Cell Neurosci.*, **3** (2010), 19

[106] Freeman J. H., Steinmetz A. B., "Neural circuitry and plasticity mechanisms underlying delay eyeblink conditioning", *Learn Mem.*, **18**:10 (2011), 666–677

[107] Yang Y., Lei C., Feng H., Sui J.-A., "The neural circuitry and molecular mechanisms underlying delay and trace eyeblink conditioning in mice", *Behav. Brain Res.*, **278** (2015), 307–314

[108] Jirenhed D.-A., Bengtsson F., Hesslow G., "Acquisition, extinction, and reacquisition of a cerebellar cortical memory trace", *J. Neurosci.*, **27**:10 (2007), 2493–2502

[109] Ito M., Kano M., "Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex", *Neurosci. Lett.*, **33**:3 (1982), 253–258

[110] Ito M., *The Cerebellum and Neural Control*, Raven, New York, 1984, 580 c.

[111] Mandwal A., Orlandi J. G., Simon C., Davidsen J., "A biochemical mechanism for time-encoding memory formation within individual synapses of Purkinje cells", *PLoS One*, **16**:5 (2021), № e0251172

[112] Delgado-Garcia J. M., Gruart A., "The role of interpositus nucleus in eyelid conditioned responses", *Cerebellum*, **1**:4 (2002), 289–308

[113] Delgadoz-Garcia J. M., Gruart A., "Firing activities of identified posterior interpositus nucleus neurons during associative learning in behaving cats", *Brain Res. Rev.*, **49**:2 (2005), 367–376

[114] Delgado-Garcia J. M., Gruart A., "Building new motor responses: eyelid conditioning revisited", *Trends Neurosci.*, **29**:6 (2006), 330–338

[115] Aksenov D., Serdyukova N., Irwin K., Bracha V., "GABA neurotransmission in the cerebellar interposed nuclei: involvement in classically conditioned eyeblinks and neuronal activity", *J. Neurophysiol.*, **91**:2 (2004), 719–727

[116] Aksenov D. P., Serdyukova N. A., Bloedel J. R., Bracha V., "Glutamate neurotransmission in the cerebellar interposed nuclei: involvement in classically conditioned eyeblinks and neuronal activity", *J. Neurophysiol.*, **93**:1 (2005), 44–52

[117] Blankenship M. R., Huckfeldt R., Steinmetz J. J., Steinmetz J. E., "The effects of amygdala lesions on hippocampal activity and classical eyeblink conditioning in rats", *Brain Res.*, **1035**:2 (2005), 120–130

[118] Weisz D. J., Harden D. G., Xiang Z., "Effects of amygdala lesions on reflex facilitation and conditioned response acquisition during nictitating membrane response conditioning in rabbit", *Behav. Neurosci.*, **106**:2 (1992), 262–273

[119] Lee T., Kim J., "Differential effects of cerebellar, amygdalar, and hippocampal lesions on classical eyeblink conditioning in rats", *J. Neurosci.*, **24**:13 (2004), 3242–3250

[120] Grishaev A.V., Sazonov V.F., "The role of active and passive properties of dendritic treein multidimensional postsynaptic potentials integration", *Usp. fiziol. nauk.*, **51**:3 (2020), 87–104 (In Russian)

[121] Sazonov V.F., Sazonov I.V., Grishaev A.V., "The concept of regimity in neuron work as a functional alternative to the structural plasticity

in the computer simulation of interneuronal interactions", *Journal Neurocomputers*, **22**:5 (2020), 43–53 (In Russian)

[122] Grishaev A.V., Sazonov V.F., "Dendritic mechanisms of regulation of activity and plasticity of neocortical pyramidal neurons", *Usp. fiziol. nauk.*, **49**:3 (2018), 104–118 (In Russian)

[123] Friston K., Kiebel S., "Cortical circuits for perceptual inference", *Neural Netw.*, **22**:8 (2009), 1093–1104

[124] Clark, A., "Whatever next? Predictive brains, situated agents, and the future of cognitive science", *Behav. Brain Sci.*, **36**:3 (2013), 181–204

[125] Gładziejewski, P., "Predictive coding and representationalism", *Synthese*, **193** (2016), 559–582

[126] Hohwy, J., *The Predictive Mind*, Oxford University Press, New York, 2013, 288 c.

[127] Hohwy, J., "The self-evidencing brain", *Noûs*, **50**:2 (2014), 259–285

[128] Kant I., *Prolegomena to any future metaphysics that can appear as a science. / / Kant I. Works in six volumes. V. 4, I. 1.*, M., 1965 (In Russian), 544 c.

[129] Clark, A., *Predicting peace: the end of the representation wars. In: Metzinger T., Windt J. M. Frankfurt am Main: MIND Group*, MIND Group, Frankfurt am Main, 2015, 1–7 c.

[130] Clark, A., "Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil", *Noûs*, **51**:4 (2016), 727–753

[131] Bruineberg, J., Kiverstein, J., Rietveld, E., "The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective", *Synthese*, **195**:6 (2018), 2417–2444

[132] Lenin V.I., *Complete works: in 55 volumes / V. I. Lenin; Institute of Marxism-Leninism under the Central Committee of the CPSU - 5th ed. V. 18. Materialism and empirio-criticism.*, State Publ. House of Polit. lit., M., 1968 (In Russian), 7—384 c.

[133] Yang X.-S., *Optimization Techniques and Applications with Examples*, John Wiley & Sons, Inc. Hoboken, New Jersey, 2018, 384 c.

[134] Friston K., FitzGerald T., Rigoli F., Schwartenbeck P., Pezzulo G., "Active Inference: A Process Theory", *Neural Comput.*, **29**:1 (2017), 1–49