**Принцип свободной энергии**

**Карл Фристон**

Professor of Neurology, University College London

*Невролог Карл Фристон о гипотезе байесовского мозга, максимизации маргинальной вероятности и минимизации свободной энергии*

ПАРТНЕР ПРОЕКТА: *Вместе с*[*Британским Советом*](http://www.britishcouncil.ru/)*в России мы подготовили проект*[*«Британские ученые»*](https://postnauka.ru/specials/british-scientists)*, посвященный истории британской науки. В этой лекции профессор неврологии Университетского колледжа Лондона Карл Фристон рассказывает о принципе организации живых систем.*

Принцип свободной энергии изначально возник в области системной нейронауки как принцип понимания того, что делает мозг и как он это делает. Впоследствии принцип оказался таким простым и действенным, что стал применяться в различных контекстах. Можно рассматривать принцип свободной энергии как принцип организации всех систем, имеющих черты живого.

Я начал с этого, поскольку существуют два пути объяснения и понимания принципа свободной энергии. Можно пойти по пути Гельмгольца, который в XIX веке пытался понять бессознательные умозаключения мозга, и основываться на анализе путем синтеза и психологии или же по пути последних впечатляющих разработок в машинном обучении, таких как, например, машина Гельмгольца, над которой работал Джеффри Хинтон. А затем это было переосмыслено в контексте энактивизма, воплощенного познания. Если обобщить эти принципы, можно прийти к принципу свободной энергии или же начать более абстрактно и задаться простыми вопросами о том, что значит быть живым. А если вы живой и существуете, какие типы поведения вы будете демонстрировать? В самом деле, ответив на эти вопросы, вы придете к тем же ответам, к которым вы пришли бы, следуя по историческому пути.

Рекомендуем по этой теме:

**Аутизм**

Ради краткости изложения начнем более абстрактно. Я буду отталкиваться от простейшего допущения, что объекты существуют, и постараюсь показать, как можно в итоге прийти к представлению о мозге как о машине вывода. Это иногда называют гипотезой байесовского мозга — мозга как одного из лучших примеров органа, который активно строит объяснения, самостоятельно отбирая информацию о мире. Этот подход энактивизма очень важен, потому что мозг не только должен объяснять всю входящую сенсорную информацию, но и решать, что из этой информации отбирать. Он сам отвечает за сбор информации, фактов для собственных предсказаний и представлений о мире. Но я ушел далеко вперед. Теперь необходимо вам разъяснить, почему существующая система будет вести себя так, будто у нее есть некоторая модель мира и она пытается собирать информацию, подтверждающую ее модель.

Прежде чем начать рассказ, следует отметить, что необходимо четко отделить предмет, о котором мы будем говорить, от всего остального. Если не провести границу между этим предметом и всем остальным, то мы не увидим разницы между ними. На статистическом языке это разделение (граница) называется «покрытие Маркова». Это просто математический способ разделения состояний некой абстрактной системы, будь то организм, культура, жизнь, клетка или мозг, на объекты внутри этой системы, которые ей принадлежат, и объекты вне ее — те, что за ее рамками. Таким образом, это может быть клетка и окружающая ее среда, это может быть фенотип, это могу быть я и мое окружение. И на любом уровне должно быть это разграничение. Само существование этого разграничения, этого покрытия Маркова в сочетании с предположением о том, что система существует некоторое время, говорит нам нечто важное о поведении внутренних состояний и состояний, которые составляют покрытие Маркова.

Это звучит немного абстрактно, но на самом деле все довольно просто. Покрытие Маркова имеет два основных аспекта. С одной стороны, есть сенсорные состояния, которые определяются тем, что они не влияют на внешние состояния, но влияют на внутренние. Так что сенсорная информация, например, будет проходить через сенсорные состояния, которые переводят информацию из внешнего мира в мой внутренний, мой мозг. С другой стороны, есть активные состояния, работающие в противоположном направлении. Они влияют на состояния внешнего мира, но не находятся под их влиянием. При этом они зависят от внутренних состояний. Если я рассмотрю себя в качестве модели моего мира, мои активные состояния будут отражать то, как я двигаюсь в данный момент, в то время как мои сенсорные состояния будут представлять собой активность моих фоторецепторов и всех тех органов чувств и чувствительных эпителиев, которые у меня есть.

Давайте на минуту забудем о понятии «покрытие Маркова» и просто подумаем, что значит для системы существовать в течение некоторого времени. Это означает, что она сопротивляется тому, чтобы быть рассеянной случайными флуктуациями. В качестве самого простого примера предположим, что я капнул каплю чернил в чашку с водой и она почти мгновенно начинает растворяться — все ее молекулы «разбегаются» по среде. И такую каплю чернил я не назову живой, потому что она растворилась. Однако, если я капну чернилами в воду и, к своему удивлению, увижу, что они вновь собрались в изначальную каплю, потом она опять немного растворилась, а затем вновь собралась, как будто она дышит или как будто время обратилось вспять, вы скажете, что есть нечто странное в этой капле чернил. Она ведет себя почти как живая, и вы быстро убеждаетесь в том, что она живая.

Рекомендуем по этой теме:

**ДНК и психогенетика**

И единственной причиной, по которой вы наделите ее этим свойством самоорганизующейся жизни, биологической самоорганизацией, будет то, что она не растворяется, и единственная причина, по которой она не растворяется, заключается в том, что все ее внутренние состояния и покрытие Маркова, которое отделяет ее от остальной воды, стремятся к ее центру. Движение молекул в системе капли чернил направлено в прямо противоположную сторону от тех сил, которые пытаются растворить ее в воде. Наверное, можно операционно или математически показать, что это течение всего-навсего движется вверх по распределению вероятностей того, где могут находиться молекулы чернил. И с точки зрения математики это распределение вероятностей идентично тому, что называется байесовской моделью представления знаний.

Я не могу посвятить этому очень много времени, но одно красивое наблюдение показывает, что определяющая динамика любой системы, которая не рассеивается со временем, состоит в том, что в среднем они будут двигаться или их состояния будут перетекать из одного в другое, чтобы максимизировать вероятность наблюдений (или наблюдаемых событий) в байесовской модели представления знаний. Это означает, что, если система существует, она будет стараться максимизировать вероятность наблюдений, то есть будет маленьким байесовским двигателем. Она будет вести себя так, будто у нее есть своя модель мира. Почему? Потому что эта система (теперь вернемся к покрытию Маркова) состоит из активных и сенсорных внутренних состояний, которые описываются покрытием Маркова. Это правило, которое говорит, что все состояния должны максимизировать вероятность наблюдений, также известную как маргинальная вероятность, обратное значение которой ограничено сверху свободной энергией — отсюда и принцип свободной энергии. Все эти состояния, включая действия, должны максимизировать маргинальную вероятность или, что то же самое, минимизировать свободную энергию. Это означает, что и действия, и ощущения во внутренних состояниях занимаются одним и тем же. Это, в свою очередь, означает, что мы можем считать, что внутренние состояния, скажем, мозга моделируют мир, потому что они максимизируют байесовскую вероятность наблюдений для меня или для модели мира.

В то же время мои действия тоже пытаются максимизировать вероятность наблюдений в пользу моей модели мира. Таким образом, если говорить простым языком, практически по определению я собираю информацию, которая максимизирует количество подтверждений моего собственного существования, и это и есть принцип свободной энергии. Это следствие любой системы, которая не рассеивается, и она выглядит так, будто должна вести себя таким образом, чтобы активно набирать информацию из среды и моделировать эту информацию в качестве модели среды, чтобы максимизировать подтверждения своего существования. И это относится ко всему: от гельмгольцевского понятия бессознательного умозаключения и до его применений в современном машинном обучении, таких как машина Гельмгольца, созданная Джеффри Хинтоном и Питером Дайаном.

Эти рассуждения можно воссоздать на множестве разных уровней, что и предоставило очень удобную концепцию для понимания того, как принцип свободной энергии согласуется с биологией, анатомией и физиологией мозга. Что нам это говорит? А то, что анатомия любой системы должна содержать модель среды, в которую эта система погружена. Это означает, что если мы живем в мире, у которого есть некая глубокая иерархическая структура, в которой присутствует действие на расстоянии, как, например, окружающий меня цвет, определяемый светом, который доходит до моих глаз почти мгновенно, или падающее тело, притягиваемое гравитацией, то тогда мой мозг должен восстановить причинно-следственную связь события. Очевидно, он это и делает.

Сам факт того, что нервные клетки с длинными редкими связями соединены друг с другом через расстояние, говорит именно о том, что причинная архитектура мира, в котором мы живем, имеет это действие на расстоянии и эту разреженную связность. Более того, иерархическая структура мира воспроизводится в нейронных структурах, которые составляют иерархии коннектома или иерархическое строение функционально специализированных зон мозга. Можно пойти дальше и задаться вопросом: если мозг — это действительно статическая модель мира, в котором он живет, можем ли мы понять самые основы организации мозга, такие как разница между каналами обработки информации «что» и «где»? Интересное наблюдение из принципа функциональной специализации состоит в том, что информация «где» перерабатывается примерно здесь, а более дорсальные пути отвечают за «что». Это может быть простым отражением того факта, что мы живем в мире, где разные вещи могут быть в разных местах. Так что мы можем статистически разделить информацию об объекте и его местоположении. Если бы мы жили во вселенной, где что-то меняет свою сущность, когда оно перемещается, мы бы не смогли этого сделать. Поэтому, просто смотря на мозг, я могу сказать, в какой вселенной мы находимся, исходя из принципа свободной энергии, при условии, что мозг становится моделью окружающей его среды.

Рекомендуем по этой теме:

**Клеточные мембраны**

Принцип свободной энергии очень удобен, как кажется мне и моим коллегам, в основном потому, что он показывает связи между предыдущими теориями. Существует множество глобальных теорий мозга, с которыми необходимо считаться. Например, принцип минимальной избыточности и максимальной эффективности, согласно которому мозг старается извлечь как можно больше информации из окружающей его среды. Существуют другие теории, объясняющие, как мы выбираем и оцениваем определенные паттерны поведения. Полезно наблюдать, как все они становятся частными случаями вариационных принципов, одним из которых является принцип свободной энергии. Это означает, что теперь можно обращаться к разным дисциплинам и наблюдать, как одна определенная концепция, будь то теоретическое понятие или эмпирическое свидетельство, связана с другим теоретическим понятием, и, по сути, можно увидеть, как они подходят к одной и той же проблеме с разных сторон. Поскольку у вас есть принципиальная схема (модель), это также позволяет вам выдвинуть конкретную гипотезу о теориях процесса, которая будет согласовываться с этим принципом.

Итак, все, что я на данный момент сказал, — это то, что в принципе любое внутреннее состояние, любое действие, которое я произвожу, любое ощущение, которое у меня возникает, должно служить цели минимизации вариационной свободной энергии или максимизации маргинальной вероятности. Как? Как можно это сделать? Как это делает мозг? Если мы знаем, какова целевая функция, если мы знаем процесс и задачи, мы затем можем изобразить это в терминах процессов. Например, я могу сказать: эта минимизация вариационной свободной энергии или максимизации байесовской вероятности наблюдений — это алгоритм восхождения на вершину или алгоритм градиентного спуска. Тогда я могу записать дифференциальное уравнение, где все нейронные состояния, физиологические состояния в мозге теперь можно описать при помощи дифференциального уравнения, беря в расчет другие состояния в мозге. И, если это уравнение верно, тогда я могу связать переменные с физиологическими процессами.

И если это сделать, можно далеко зайти в понимании не только анатомии, но и физиологии. И можно сформулировать новые вопросы, поскольку существуют альтернативные процессы, которые неподвластны этому принципу, и также мозг использует техники сэмплинга (отбора проб), чтобы максимизировать вероятность наблюдений, или использует алгоритмы оптимизации восхождения на холм, вариационные схемы. Таким образом, вы начинаете формулировать множество гипотез, относящихся к теории процессов, которые можно протестировать, и они полностью согласуются с общим принципом.

**Карл Фристон**

Professor of Neurology, University College London

Начало формы