М.А. Сущин

Байесовский подход к восприятию и вопрос о реализме[1]

Введение

Настоящая статья преследует две основные цели: представить российским ученым-когнитивистам, а также специалистам в области теории познания, философии сознания и философии науки активно набирающую популярность амбициозную и претендующую на роль единой теории в современных исследованиях познания программу исследований, известную в англоязычной литературе преимущественно как «предсказывающая обработка», или «предсказывающее кодирование» (англ. predictive processing/predictive coding). Будучи, в сущности, развитием гельмгольцевского подхода к восприятию как к производимым мозгом «бессознательным умозаключениям», данная теория на первый взгляд как будто бы способствует антиреалистическому пониманию перцептивных процессов и познания в целом – недаром за идущей от Гельмгольца традицией в исследованиях восприятия закрепилось клеймо «конструктивизм»[2]. Поэтому вторая цель настоящей работы заключается в том, чтобы исследовать, действительно ли данная, как представляется, многообещающая теория в эпистемологическом плане способствует развитию антиреалистических подходов к познанию.

Однако вначале следует сказать о предыстории рассматриваемого направления. Так, хотя в своем современном облачении парадигма «предсказывающего кодирования» является продуктом синтеза весьма богатого множества идей и концепций современной математики, компьютерной науки, нейронауки и когнитивной науки, общая идея, стоящая за этим направлением, имеет долгую историю. Например, еще арабский ученый-энциклопедист XI века Ибн аль-Хайсам (Альхазен) придерживался точки зрения, что

[1] Работа выполнена при поддержке гранта РНФ, проект № 15-18-10013 «Социо-антропологические измерения конвергентных технологий».

[2] *Palmer S.* Vision Science: Photons to Phenomenology. Cambridge, MA, 1999. P. 56–57.

«многие видимые свойства воспринимаются посредством суждения и вывода»[1]. Кант, постулировавший априорные структуры (априорные формы пространства и времени и априорные категории рассудка) для оформления предоставляемого чувственностью материала[2], также иногда рассматривается в качестве одной из предтеч данного направления. Впрочем, как уже было сказано, едва ли приверженцы теории предсказывающей обработки могут быть обязаны кому-либо более, нежели знаменитому немецкому физику и физиологу Герману фон Гельмгольцу[3].

Именно Гельмгольцу принадлежит ключевая для этого направления идея, а именно что восприятие заключает в себе осуществление (вероятностных) выводов по эффектам воздействия внешнего мира на наш сенсорный аппарат о причинах этого воздействия, производимое мозгом и нервной системой без участия сознания. «Когда те нервные механизмы, окончания которых находятся с правой стороны сетчатки обоих глаз, были стимулированы, – писал Гельмгольц, – нашим обычным опытом, повторявшимся миллионы раз на протяжении всей жизни, было то, что прямо перед нами с левой стороны [как будто бы – М.С.] находился светящийся объект…. Таким образом, хотя в этих случаях не было представлено специфических сознательных выводов, все же сущностное и подлинное условие (office) для вывода было исполнено, и результат его был достигнут; попросту, конечно же, посредством бессознательных процессов ассоциации идей, происходящих на темных задворках (background) нашей памяти. Таким образом, также его результаты воздействуют на наше сознание, так сказать, как если бы нас ограничивала внешняя сила, над которой наша воля не имеет власти»[4].

---

[1] Цит. по: *Hohwy J.* The Predictive Mind. New York, 2013. P. 5.

[2] *Кант И.* Критика чистого разума. М., 1994.

[3] *Helmholtz H. von.* Concerning the Perceptions in General // Helmholtz H. von. Treatise on physiological optics. Vol. 3. Ch. 26. 3rd edn. Translated by J.P.C. Southall, 1925, *Op. Soc. Am.* Section 26, reprinted New York, 1962. P. 1–37.

[4] *Helmholtz H. von.* Op. cit. P. 26. Со своей стороны мы не можем не отметить весьма примечательное сходство между приведенной выше формулой Гельмгольца о «бессознательных умозаключениях» и основными идеями возникшей через сто лет вычислительной когнитивной науки. Правда, здесь необходимо соблюдать осторожность, поскольку Гельмгольц связывал бессознательные перцептивные выводы с деятельностью нервной системы, тогда как ранние когнитивисты придерживались так называемого принципа «множественной реализованности» (англ. multiple realizability, multiple instantiability), направленного на установление автономии когнитивных процессов от конкретных физических параметров систем, реализующих эти процессы (и, как следствие, на установление автономии когнитивной психологии от нейронауки).

Также, по аналогии со значением экспериментов в науке Гельмгольц в весьма современном ключе подчеркивал роль и необходимость действия в процессе восприятия внешнего мира. Именно действия и наша произвольная исследовательская двигательная активность, утверждал он, играют неоценимую роль в усилении корректности суждений о причинах наших восприятий. И если бы кто-либо был в состоянии сделать так, чтобы только объекты внешнего мира проходили перед нашим взором «без того, чтобы мы могли что-либо с ними сделать, вероятно, мы никогда бы не смогли найти свой путь среди подобной оптической фантасмагории»[1].

В XX веке традиция, заложенная Гельмгольцем, была продолжена в работах в таких известных исследователей восприятия, как Ричард Грегори[2], Ирвин Рок[3], Гаетано Канизза[4]. В сходном ключе Джером Брунер в рамках «психологии нового взгляда»[5], Ульрик Найссер в классической работе «Когнитивная психология»[6] и многие другие исследователи подчеркивали роль гипотез, категоризации, внимания, ожиданий, памяти и знания в оформлении текущего сенсорного входа.

Другой лейтмотив теории «предсказывающей обработки» (далее сокращенно – ТПО) – идея минимизации (коррекции) в ошибок в предсказании (англ. prediction error), которые неизбежно возникают, когда, как постулируется, генерируемые мозгом гипотезы сопоставляются с текущими сенсорными сигналами – имея определенные корни в классической кибернетике, в значительной степени концептуально навеян современными техниками сжатия данных в процессе передачи и обработки сигналов, разработанными инженерами Лаборатории Белла в 1950-х гг. Как разъясняет философ Энди Кларк: «рассмотрим базовую задачу, такую, как передача изображения: для большинства изображений значение одного пикселя аккуратно (regularly) предсказывает значение ближайших

[1] *Helmholtz H. von.* Op. cit. P. 31.

[2] *Gregory R. L.* Perceptions as Hypotheses // Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 1980. Vol. 290. No. 1038. P. 181–197. *Gregory R. L.* Knowledge in perception and illusion // Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 1997. Vol. 352. No. 1358. P. 1121–1128.

[3] *Rock I.* The Logic of Perception. Cambridge, MA, 1983.

[4] *Kanizsa G.* Seeing and thinking // Acta Psychologica. 1985. Vol. 59. No. 1. P. 23–33.

[5] *Брунер Дж.* О готовности к восприятию // Брунер Дж. Психология познания. М., 1977. С. 13–64.

[6] *Neisser U.* Cognitive Psychology. New York and London, 2014.

к нему соседних [пикселей], с различиями, отмечающими важные свойства, такие, как границы между объектами. Это означает, что код для большого изображения может быть сжат (для надлежащим образом информированного получателя) посредством кодирования только «непредвиденного» изменения: случаев, где действительное значение отличается от предсказанного. То, что должно быть передано, следовательно, есть просто разница (также известная, как «ошибка в предсказании») между действительным текущим сигналом и предсказанным сигналом…. Наследники данной техники сжатия данных в настоящее время используются в JPEG, в различных формах сжатия аудио без потерь и в сжимающем движения кодировании для видео»[1].

Именно в честь Гельмгольца получил свое название новый тип нейронных сетей, так называемые «Машины Гельмгольца»[2], существенная особенность обучения и работы которых позволяет рассматривать их в контексте становления основных идей ТПО. Так, например, в отличие от более ранних коннекционистских алгоритмов и процедур обучения (таких, как метод обратного распространения ошибки[3], который имел ограниченные ресурсы для работы в глубоких, многослойных сетях и в процессе обучения требовал большое количество заранее помеченных данных, которые не всегда имеются под рукой), данный вид сетей основывался на принципиально новом подходе к обучению. Его суть, обобщенно говоря, заключалась в том, что для выполнения задач распознавания и классификации изображений (например, написанных от руки цифр) нейронная сеть выучивала модель (так называемую генеративную модель), которая сама непосредственно производила сенсорные данные (изображения) таким образом, чтобы они в максимальной степени соответствовали образцу (тех же цифр), который и подлежал распознаванию.

Каким образом сеть выучивала подходящую генеративную модель, если для этого нужно было использовать обученные восходящие пути распознавания, для использования которых в свою очередь уже необходимо было располагать соответствующим образом настроенными нисходящими генеративными путями? Чтобы

[1] *Clark A.* Whatever next? Predictive brains, situated agents, and the future of cognitive science // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 182–183.

[2] *Dayan P., Hinton G. E., Neal R. M., Zemel R. S.* The Helmholtz Machine // Neural Computation. 1995. Vol. 7. No. 5. P. 889–904.

[3] Классический метод обучения, при котором разница между текущим выходным состоянием сети и желаемым выходным состоянием распространяется в обратном от выходного слоя направлении с использованием тех же самых весов соединений, которые использовались для прямой связи в сети.

обойти «проблему курицы и яйца» ученые использовали новый самонастраиваемый алгоритм «бодрствование-сон»[1]: «Оказывается, – объясняет ведущий специалист в области обучения нейронных сетей Джеффри Хинтон, – что мы можем выучить оба набора весов (сила соединения между двумя искусственными нейронами – М.С.), начиная с небольших случайных значений и чередуясь между двумя фазами обучения. В фазе «бодрствование» весы распознавания используются, чтобы стимулировать (drive) узлы снизу вверх, и бинарные состояния узлов в прилегающих слоях могут затем использоваться, чтобы обучать генеративные весы. В фазе «сон» нисходящие генеративные соединения используются, чтобы управлять сетью, чтобы она продуцировала фантазии из своей генеративной модели»[2]. Таким образом, дополнительно настроенная при помощи процедуры, именуемой «пре-обучением» слой за слоем (англ. pre-training) нейронная сеть с генеративной моделью из трех слоев детекторов свойств могла правильно распознать множество плохо написанных от руки цифр (проще говоря, каракуль), в частности, таких, которые никогда не попадались ей ранее[3].

Приблизительно в одно время с развитием генеративных методов в обучении искусственных нейронных сетей исследователи из университетов США и Великобритании, работавшие на стыке математики и психологии, опираясь на байесовскую теорию принятия решений и методы математической статистики, начали разрабатывать детальные байесовские модели восприятия и перцептивных процессов[4]. В настоящее время амбиции этой теории растянулись таким образом, что она стремится выступить в качестве единой теории для объяснения не только восприятия, но и столь

[1] *Hinton G. E., Dayan P., Frey B. J., Neal R. M.* The «wake-sleep» algorithm for unsupervised neural networks // Science. 1995. Vol. 268. No. 5214. P. 1158–1161. Использование данного алгоритма, разумеется, не было ограничено только «Машинам Гельмгольца», он также использовался для обучения сетей, именуемых «Ограниченными машинами Больцмана» – см.: *Hinton G. E.* Learning multiple layers of representation // Trends in Cognitive Sciences. 2007. Vol. 11. No. 10. P. 428–434.

[2] *Hinton G. E.* To Recognize Shapes, First Learn to Generate Images [Электронный ресурс]. URL: http://www.cs.toronto.edu/~hinton/absps/montrealTR.pdf (дата обращения: 05.04.2016).

[3] Весьма впечатляющие демонстрации работы сетей, натренированных посредством упомянутого выше алгоритма «бодрствование-сон» и процедуры «пре-обучения», могут быть найдены на личной странице Джеффри Хинтона: http://www.cs.toronto.edu/~hinton/adi/index.htm

[4] *Rescorla M.* Bayesian Perceptual Psychology [Электронный ресурс]. URL: http://www.philosophy.ucsb.edu/docs/faculty/papers/bayesian.pdf (дата обращения: 07.04.2016). Perception as Bayesian Inference / Edited by D. C. Knill, W. Richards. New York: Cambridge University Press, 1996.

разнообразных когнитивных, психофизических и ментальных феноменов, как действие[1], внимание, иллюзии, психические расстройства[2], а также в некоторой степени – опыт, сознание, Я[3] и эмоции[4]. Так, к примеру, постулируемый одним из лидеров этого направления известным нейроученым Карлом Фристоном «Принцип свободной энергии»[5] пытается объединить под своими знаменами столь разнообразные теории, как, собственно, ТПО и гипотезу байесовского мозга, классическую теорию клеточных ансамблей Дональда Хебба и теорию нейродарвинизма Джеральда Эдельмана, теорию оптимального контроля и теорию внимания как направленного соревнования. Наконец, Энди Кларк, чья статья в журнале «Науки о поведении и мозге» вызвала оживленную дискуссию ряда известных ученых и философов и внесла неоценимый вклад в популяризацию идей этого направления, высказал смелое предположение, что данная теория весьма близка к преодолению трудностей, ранее препятствовавших построению единой теории разума, мозга и действия[6].

Данная статья, тем не менее, будет по преимуществу ограничена вопросами ТПО и применения байесовских конструктов к описанию процессов восприятия. Классическими оказавшими наибольшее влияние на развитие ТПО считаются работы знаменитого математика Дэвида Мамфорда, работа когнитивистов Рао и Балларда, а также, собственно, исследования Карла Фристона и его сторонников. Именно на работах этих исследователей и будет сосредоточено внимание в основной части нашей работы.

Затем мы перейдем к рассмотрению вопроса об отношении ТПО к реализму и реалистически направленным концепциям восприятия. Последние, как известно, в значительной степени ассо-

---

[1] См., например: *Friston K.* What Is Optimal About Motor Control // Neuron. Vol. 72. No. 3. P. 488–498. *Rescorla M.* Bayesian Sensorimotor Psychology [Электронный ресурс]. URL: http://www.philosophy.ucsb.edu/docs/faculty/michael-rescorla/rescorla_bayesian-sensorimotor-psychology.pdf (дата обращения: 09.04.2016).

[2] *Fletcher P. C., Frith C. D.* Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia // Nature Reviews Neuroscience. 2009. Vol. 10. No. 1. P. 48–58.

[3] *Hohwy J.* Op. cit.

[4] *Seth A.* The Cybernetic Bayesian Brain [Электронный ресурс]. URL: http://open-mind.net/papers/the-cybernetic-bayesian-brain (дата обращения: 21.07.2016).

[5] *Friston K.* The free-energy-princple: a unified brain theory? // Nature Review Neuroscience. 2010. Vol. 11. No. 2. P. 127–138.

[6] *Clark A.* Op. cit. P. 200.

циируются с именем Дж. Дж. Гибсона, экологическая теория которого считается едва ли не главным антиподом гельмгольцевского вычислительного подхода. Теория Гибсона в числе прочего также оказала непосредственное влияние на формирование другого влиятельного направления в современных исследованиях познания – так называемого «ситуативного/воплощенного познания», оспорившего роль и влияние более ранних вычислительных подходов в когнитивной науке. С появлением на горизонте новой масштабной теории многие исследователи начали проводить сравнения или же искать возможные точки соприкосновения между двумя теориями. Поэтому вопрос об отношении ТПО к реализму является в некотором смысле и вопросом об отношении этого недавно возникшего направления к более ранним идеям «ситуативного/воплощенного познания», и в части, касающейся эпистемологической оценки ТПО, мы неизбежно будем вынуждены хотя бы вкратце затронуть вопрос об отношении двух идей. Приступим к рассмотрению.

Предсказывающая обработка: первый взгляд

Итак, каковы основные положения ТПО? Пожалуй, один из лучших способов начать знакомство с идеями байесовской психологии восприятия и парадигмы «предсказывающего кодирования» предоставляют работы такого известного математика современности, как Дэвид Мамфорд. Мамфорд, чьи научные интересы простираются от алгебраической геометрии до исследования вычислительных основ работы головного мозга, изложил свои нейрокогнитивистские взгляды в серии статей, главными из которых считаются вторая часть работы о вычислительной архитектуре новой коры[1], касающаяся роли и взаимодействия кортико-кортикальных петлей, а также статья 2003 года об иерархическом устройстве обработки информации в зрительной коре, написанная в соавторстве со специалистом в области компьютерной науки Тай Сингом Ли[2].

В изложенной в двух частях классической работе 1991-1992 гг. Мамфорд, основываясь на известных нейроанатомических данных о принципиальном единообразии организации новой коры головного мозга млекопитающих, выдвинул предположение, что таким образом новая кора может воплощать в себе некоторые универ-

[1] *Mumford D.* On the computational architecture of the neocortex: II the role of cortico-cortical loops // Biological Cybernetics. 1992. Vol. 66. No. 3. P. 241–251.

[2] *Lee T. S., Mumford D.* Hierarchical Bayesian inference in the visual cortex // Journal of the Optical Society of America. 2003. Vol. 20. No. 7. P. 1334–1348.

сальные принципы функционирования и что эти принципы могут быть поняты при помощи концептуальных средств ряда вычислительных теорий (таких, например, как теории распознавания паттернов). В первой части работы[1], которая не представляет для нас здесь значительного интереса, исследовалась вычислительная роль таламо-кортикального комплекса, в котором таламус, по предположению Мамфорда, играет роль своего рода «активной рабочей области» (англ. «active blackboard»), синтезируя и в динамической манере заново представляя для коры текущие сенсорные данные, сообщенные ему обратно после некоторой первичной обработки в коре. Во второй, более известной и цитируемой части работы Мамфорд предложил теорию возможных иерархических вычислительных отношений между высшими и низшими регионами непосредственно коры мозга. Данную теорию, исходя из использованной Мамфордом терминологии, можно условно обозначить как теорию «шаблон/остаток» (англ. «template/residual»).

Смысл подхода Мамфорда достаточно прост (однако вместе с тем его уяснение чрезвычайно важно для уяснения смысла всего рассматриваемого нами направления): иерархически более высокий регион коры *A*, генерирующий и содержащий в себе определенные ожидания и предсказания, которые можно назвать шаблонами, относительно того, что будет зарегистрировано сенсорным входом (например, лицо, дом и т.п.; соответственно, имеются в виду шаблоны лица, дома и т.п.), связан с более низкоуровневым регионом *B*, который непосредственно имеет дело с этой самой более конкретной сенсорной информацией. Далее, согласно Дэвиду Мамфорду, общее взаимодействие и взаимная динамика слоев *A* и *B* может быть описана посредством следующей схемы: (1) слой *A* отправляет сигнал слою *B*, содержащий предсказание, или шаблон, для каждого воспринимаемого объекта *O*; (2) затем сигнал переводится в понятную слою *B* форму (например, высокоуровневый шаблон лица в специфические паттерны линий, форм, границ и цветов и т.п.), и этот нижестоящий слой сопоставляет полученные шаблоны с тем, что ему известно о текущем сенсорном входе, и, если предсказания области *A* не являются исчерпывающими, вычисляет «остаток, описание той части мира, которая не ожидалась или предсказывалась»[2]; (3) соответственно, эта разница, или остаток, между тем, что предсказано, и тем, что воспринято, посылается обратно в вышестоящий регион *A*, который (4) «… модифици-

[1] *Mumford D.* On the computational architecture of the neocortex: I The role of the thalamo-cortical loop // Biological Cybernetics. 1991. Vol. 65. No. 2. P. 135–145.

[2] *Mumford D.* On the computational architecture of the neocortex: II the role of cortico-cortical loops // Biological Cybernetics. 1992. Vol. 66. No. 3. P. 247.

рует свои параметры в гибком шаблоне, чтобы попытаться улучшить соответствие, и посылает его обратно *B*»[1]. Наконец, что наиболее важно, после нескольких циклов, полагает Мамфорд, «… либо найдено хорошее соответствие, и остаток является приемлемо малым, либо гипотеза отвергается, и зона *A* возвращается к другой ранее сдерживавшейся гипотезе»[2].

Таким образом, в наиболее стабильном состоянии, заключает Мамфорд, зона *A* дожна генерировать сигнал, который с учетом определенного уровня шума будет почти совершенно предсказывать, с какой именно сенсорной информацией в данный момент имеет дело регион *B*, и нейроны зоны *B* (о предложенных Мамфордом нейронных коррелятах чуть позже), ответственные за вычисление остатка, не будут разряжаться вообще.

А что происходит в противоположном случае, когда мы, например, обнаруживаем себя в незнакомой обстановке или же крайне озадачены чем-либо и не имеем соответствующих ожиданий? «В другой крайности, если вы просыпаетесь в незнакомом месте без ожиданий или же всецело удивлены чем-либо, – пишет Дэвид Мамфорд, – то алгоритм стартует с чистого листа в зоне *A*. Затем *B* просто отправляет *A* всю свою картину мира, которая возбуждает некоторые возможные высокоуровневые объекты. На каждом этапе *A* выписывает на своей доске (blackboard) свои лучшие догадки на своем собственном языке (объекты и их параметры) о характере высокоуровневых объектов, найденных в картине *B*»[3].

Также Мамфордом были предложены возможные нейронные структуры, ответственные за реализацию данного алгоритма. С его точки зрения, глубокие пирамидальные клетки в слое V высокоуровневых регионов коры, оканчивающиеся в слоях I и VI нижестоящего региона, могут быть связаны с нисходящей трансляцией шаблонов, тогда как поверхностные пирамидальные клетки слоев II и III, оканчивающиеся в слое IV более высокого региона, должны участвовать в обратном сообщении остатка (ошибки в предсказании, если воспользоваться более современными терминами) в случае, если предсказание не исчерпывает сенсорный сигнал.

Таким образом, уже сейчас мы можем зафиксировать несколько ключевых элементов, характерных для стратегии ТПО в изучении восприятия (и познания в целом), а именно что в мозге существуют определенные высокоуровневые структуры, которые можно называть предсказаниями или априорными ожиданиями (англ. prior

[1] Ibid.
[2] Ibid.
[3] Ibid.

expectations)[1] и которые в иерархической манере предвосхищают (или, по крайней мере, стремятся предвосхитить) то, что мозг может воспринять в следующий момент. А также что, вопреки классическим представлениям, в мозге во время восприятия имеет место своего рода инверсия между сенсорными путями прямой и обратной связи, поскольку нисходящие пути обратной связи, с этой точки зрения, определяя или предвосхищая содержание сенсорного опыта, в функциональном отношении выступают как пути прямой связи, а анатомические пути прямой связи, сообщая наверх разницу между ожиданием и текущим восприятием, в функциональном отношении предстают как пути обратной связи на нисходящую модель мира.

Другая классическая модель инверсии прямой и обратной сенсорной связи в мозге, в ряде аспектов развивающая и детализирующая темы, обозначенные в работе Дэвида Мамфорда, была предложена в совместной работе[2] когнитивистов Рао и Балларда, представивших «предсказывающую» интерпретацию так называемых экстраклассических феноменов рецептивных полей (характерных для определенных нейронов зрительных кортикальных областей V1, V2, V4 и MT). Смысл этого феномена, известного также как «подавление краев» стимулов (англ. endstopping, end-inhibition), заключается в том, что нейроны, типичным образом разряжающиеся в присутствии определенного стимула (например, линии определенной длины и ориентации), перестают разряжаться, если такого же рода стимул простирается за пределы их классического рецептивного поля[3] (т.е., например, в присутствии аналогичной линии, однако чуть большей длины). Почему? Рао и Баллард полагают, что данный эффект может быть интерпретирован с позиции модели иерархического предсказывающего кодирования, согласно которой зрительные кортикальные нейроны с экстраклассическими эффектами РП являются детекторами ошибки, сообщающей наверх «… разницу между входным сигналом и его статистическим предсказанием, основанным на эффективной внутренней модели естественных изображений [содержащейся на

[1] В английском языке для описания этих структур в последнее время принято использовать краткое слово prior, которое является сокращением от одного ключевых компонентов теоремы Байеса, так называемой априорной вероятности (англ. prior probability).

[2] *Rao R. P. N., Ballard D. H.* Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects // Nature neuroscience. 1999. Vol. 2. No. 1. P. 79–87.

[3] Определенная область сенсорного пространства (например, зрительного поля или тела агента), стимуляция которой приводит к возбуждению нейрона.

«высших этажах» вычислительной кортикальной иерархии – М.С.]»[1].

В поддержку своей гипотезы Рао и Баллард реализовали несколько вычислительных моделей с использованием специальных натренированных на естественных изображениях нейронных сетей, в которой каждый уровень «пытается предсказать реакции следующего уровня снизу через соединения обратной связи…. Ошибка между этим предсказанием и действительной реакцией затем отсылается обратно посредством соединений прямой связи. Этот ошибочный сигнал затем используется, чтобы скорректировать оценку входного сигнала на каждом уровне… Нижний уровень оперирует на небольших пространственных (и, возможно, временных) шкалах, тогда как верхние уровни оценивают свойства сигналов на больш**и**х шкалах, поскольку высокоуровневый модуль предсказывает и оценивает реакции нескольких низкоуровневых модулей»[2].

Так почему же, согласно интерпретации Рао и Балларда, нейроны, разряжавшиеся в присутствии определенного стимула, переставали разряжаться, если подобный стимул расширялся за пределы их классического РП? Дело в том, что в большинстве естественных изображений, в том числе тех, которые использовались исследователями для обучения нейронных сетей, небольшие по размерам стимулы (например, короткие линии) редко являются изолированными от других объектов зрительной сцены. Таким образом, «верхние этажи» вычислительной иерархии выучивают на основе таких статистически преобладающих изображений модель, которая далее генерирует предсказания, пошагово спускающиеся вниз по вычислительной иерархии через соединения обратной связи вплоть до последнего слоя (к примеру, сетчатки), имеющего дело непосредственно с воздействием внешней среды. В случае ошибочности предсказаний (если встречается стимул, не выходящий за пределы их классического РП), соответственно, возбуждаются нейроны-детекторы ошибок на «нижних этажах» иерархии, сообщающие об ошибке в предсказании наверх по путям прямой связи.

Именно этот вывод и подтвердили компьютерные модели Рао и Балларда, где изображения, отличающихся от средних статистически, приводили к возбуждению слоев, ответственных за обнаружение отклонения от предсказаний, генерируемых высокоуровневой моделью. Более того, оказалось, что устранение обратной связи от верхних уровней нейронной сети (которая, соответственно, лишалась своей предсказывающей силы) приводило к увеличению ак-

[1] Ibid. P. 79.
[2] Ibid. P. 80.

тивности детекторов ошибок на «нижних этажах» вычислительной иерархии на 82%[1].

Как описанная выше схема механизма минимизации ошибки в предсказании соотносится с идеей байесовского восприятия, опирающейся на знаменитую теорему Байеса как наиболее вероятную формальную модель перцептивных процессов? Для того чтобы попытаться ответить на этот вопрос, нам необходимо вкратце описать непосредственно саму теорему Байеса, или, как ее также иногда называют, правило Байеса, которое является строгим формальным методом вычисления условных вероятностей (к примеру, оценки вероятности определенной гипотезы (допустим, восприятия объекта как выпуклого или вогнутого) при наличии определенных данных (например, определенного положения источника света, уровня освещенности и т.д.)). Тогда теорема Байеса гласит, что:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)},$$

где $p(\theta|x)$ является апостериорной вероятностью гипотезы θ при условии определенных данных x; $p(x|\theta)$ есть вероятность определенных данных (или переменных) x при условии истинности гипотезы (параметров) θ (этот компонент известен также как функция правдоподобия); $p(\theta)$ является априорной, т.е. не зависящей от текущих наблюдений вероятностью гипотезы θ и $p(x)$ является вероятностью определенных данных x. Вероятность данных $p(x)$ также называется маргинальной вероятностью и часто не учитывается, поскольку, если мы, к примеру, хотим сравнить апостериорные вероятности двух гипотез (например, гипотез о причине болезни при наличии определенных симптомов), то каково бы ни было значение $p(x)$, оно не будет влиять на соотношение данных апостериорных вероятностей (т.е. сколько бы ни было в данный момент в обществе индивидов с высокой температурой $p(x)$, вероятность гриппа (гипотеза 1) при высокой температуре все равно будет на определенное неизменное значение выше вероятности лихорадки (гипотеза 2)).

Так почему же все-таки вероятность определенной гипотезы может быть оценена выше, чем вероятность некоторых других гипотез? Согласно теореме Байеса, различие апостериорных вероятностей *N* гипотез происходит от нашего предшествующего знания, которое в данном случае обозначается как априорная вероятность

[1] Ibid. P. 83.

этих гипотез, и вероятности данных при условии гипотезы. Обладая предшествующим знанием о вероятности какого-либо события или положения дел, мы можем рассудить, насколько оно в принципе возможно, отсекая (при прочих равных условиях) заведомо маловероятные гипотезы положения вещей (например, той же самой лихорадки при высокой температуре в условиях сезонной эпидемии гриппа).

Итак, теперь мы подошли к главному интересующему нас вопросу, а именно, как данная схема, как постулируется сторонниками байесовской теории, может использоваться мозгом в процессе восприятия (прежде всего, конечно, внешнего мира, хотя, как отмечалось выше, некоторые авторы утверждают, что байесовский подход может быть расширен и для анализа интероцепции). Весьма удобным для иллюстрации работы этого подхода является в некоторой степени искусственно сконструированный, однако, тем не менее, давно известный феномен так называемого бинокулярного соперничества (англ. binocular rivalry). Данный феномен был впервые описан в 1593 году итальянским врачом и философом Джамбаттиста Делла Порта и заключается в следующем. Предположим, что мозг сталкивается с задачей анализа сложной зрительной сцены, состоящей из аккуратно наложенных друг на друга полупрозрачных изображений дома и лица. В этой ситуации мозг, вопреки интуитивным ожиданиям, не будет воспринимать данное изображение как составное изображение полупрозрачных наложенных друга на друга изображений дома и лица, а будет попеременно всецело фокусироваться на изображении либо дома, либо лица (даже с учетом того факта, что они являются полупрозрачными), игнорируя другую часть изображения. Бинокулярное соперничество, как и любой другой пример так называемого бистабильного восприятия (куб Неккера, лестница Шредера, монокулярное соперничество), интересно для нас еще и тем, что мозг в данном случае имеет дело с равновероятностными интерпретациями зрительной сцены, к которым он попеременно обращается.

Попытка объяснения данного феномена с позиции байесовского предсказывающего подхода была представлена в совместной работе Якоба Хохви, Андреаса Ропстерффа и Карла Фристона[1]. Для объяснения феномена бинокулярного соперничества, подчеркивают авторы, необходимо прежде всего объяснить две его ключевые особенности, а именно (1) почему мозг в каждый конкретный момент времени предпочитает исключительно одну интерпре-

[1] *Hohwy J., Roepstorff A., Friston K.* Predictive coding explains binocular rivalry: An epistemological review // Cognition. 2008. Vol. 108. No. 3. P. 687–701.

тацию зрительной сцены, а не другую (скажем, лицо, а не дом) или, что наиболее важно, не своеобразное смешение двух объектов в один объект (дом-лицо), которое в действительности и присутствует на доступном зрительной системе изображении; (2) почему мозг попеременно «переключается» между различными интерпретациями зрительной сцены, а не «застывает» на какой-либо одной выбранной гипотезе[1].

В терминах привлекаемой авторами предсказывающей байесовской модели первая проблема получает решение следующим образом. Предположим, что гипотезы только дома θ_h или только лица θ_f одинаково соотносятся с представленным неопределенным стимулом. Предположим, далее, что гипотеза двойного объекта дом-лицо θ_{mix} лучше соотносится с вероятностью текущих данных $p(x|\theta_{mix})$ (поскольку эта гипотеза, собственно, и соответствует тому, что в действительности представлено на зрительной сцене, а именно некой смеси дома и лица). Тем не менее, мозг в каждый момент останавливается только на одной интерпретации зрительной сцены, исключая гипотезу смешанного объекта. В условиях равной вероятности данных при условии гипотез дома и лица выбор между ними может быть осуществлен только на основе априорной вероятности, т.е. предшествующего знания, свидетельствующего, что один объект более вероятен, чем другой. Руководствуясь априорной вероятностью, мозг, таким образом, «заключает», что наиболее вероятная гипотеза состоит в том, что перед ним в данный момент на зрительной сцене, к примеру, находится лицо, а не дом[2] (и не соединенные в единый объект дом и лицо).

А почему с этой точки зрения мозг практически полностью исключает гипотезу смешанного объекта даже с учетом того, что она, как было сказано, лучше соотносится с тем, что представлено на изображении (т.е. текущими данными), нежели какая-либо из гипотез единичных объектов? Ответ байесовской теории аналогичен данному ранее: благодаря крайне низкой априорной вероятности встречи в большинстве типичных ситуаций такого причудливого объекта, как дом-лицо[3], мозг заведомо отсекает возможность его

[1] Ibid. P. 690.

[2] Или наоборот. Вопрос о происхождении априорных ожиданий будет рассмотрен ниже.

[3] Нужно особо подчеркнуть, что речь идет здесь о восприятии невероятного объекта, смешения дома и лица, а не, скажем, о типичном, происходящем много раз в течение дня восприятии лица на фоне дома (или, что менее типично, дома на фоне лица). Мозг в данном случае руководствуется крайне низкой априорной вероятностью пространственно-временного сосуществования дома и лица, соот-

восприятия, выбирая далее из оставшихся гипотез с более высокой априорной вероятностью (т.е. гипотез единичных объектов, поскольку вероятность *a priori* встретить подобного рода объекты в типичной обстановке несоизмеримо выше). Наконец, выбранная гипотеза (в байесовских терминах, гипотеза с наибольшей апостериорной вероятностью) и осуществляет контроль нисходящей динамики процесса восприятия, определяя, как постулируется, его содержание и подавляя предсказываемые этой гипотезой восходящие сенсорные сигналы.

Впрочем, как известно, выбрав определенную интерпретацию зрительной сцены, мозг не задерживается на ней перманентно, а вскоре «переключается» с данной гипотезы на другую и затем обратно и т.д. Как байесовская теория объясняет подобное своеобразное чередование ведущих гипотез? Согласно Хохви, Ропстерфу и Фристону, характерное для бинокулярного соперничества чередование интерпретаций зрительных сцен в мозге происходит в силу того достаточно очевидного обстоятельства, что господствующая в данный момент гипотеза может предсказать (и объяснить) только одну часть доступной зрительной сцены (скажем, полупрозрачное изображение лица), оставляя другую (изображение дома) незатронутой. Поэтому сигналы, ассоциируемые с незатронутой господствующей гипотезой частью изображения, составляют сообщаемую на «верхние этажи» вычислительной иерархии ошибку в предсказании, с целью минимизации которой мозг должен найти ей адекватное объяснение. Таким образом, мозг вынужден обратиться к альтернативной интерпретации, которой первоначально была присвоена менее высокая апостериорная вероятность. Как нетрудно догадаться, вскоре схожий сценарий повторяется и для этой гипотезы, и мозг, как и при восприятии куба Неккера, оказывается в своего рода бистабильном перцептивном режиме, чередуя между собой конкурирующие интерпретации видимых сцен и изображений.

По словам философа Энди Кларка, данная байесовская интерпретация «… является глубоким объяснением существования конкуренции между определенными высокоуровневыми гипотезами в первую очередь. Они соперничают, – пишет Кларк, – потому что система выучила, что «только один объект может существовать в

ветственно, присваивая данной гипотезе низкую итоговую апостериорную вероятность. Правда, все же, как отмечает Энди Кларк, в случае бинокулярного противоречия такие смешанные перцепты действительно иногда имеют место быть, но лишь на определенные краткие мгновения (особенно при первых взглядах на изображение), практически сразу же уступая место более вероятным претендентам на объяснение зрительной сцены.

одном и том же месте в одно и то же время» (Хохви и др. 2008, с. 691). (Однако здесь, очевидно, требуется осторожность, поскольку отдельное состояние мира может быть последовательно схвачено посредством многих высокоуровневых сюжетов, которые не нуждаются в соперничестве аналогичным образом: например, восприятие картины как значимой, как произведения Рембрандта, как изображения коровы и т.д.)»[1].

Конечно, бинокулярное противоречие и другие примеры бистабильного восприятия крайне редко встречаются в большинстве ситуаций, с которыми имеет дело наша перцептивная система. Однако формат настоящей статьи не позволяет нам вдаваться в детальное обсуждение иных иллюстраций работы байесовского механизма, равно как и возможностей его приложения к некоторым другим упомянутым выше проблемам когнитивной науки (проблемам Я, сознания, эмоций, психических расстройств и т.д.). Основные проблемы и затруднения байесовского подхода к восприятию и познанию получат краткое освещение в заключении данной работы. Сейчас нам необходимо внести определенные завершающие штрихи, которые позволят увидеть представленную выше картину механизма минимизации ошибки в предсказании в рамках возможной более общей перспективы «принципа свободной энергии», предлагаемой британским нейроученым Карлом Фристоном.

Минимизация свободной энергии

Наш рассказ о байесовском подходе и ТПО был бы неполным, если бы мы не упомянули о недавней попытке переформулировать их основные положения в рамках потенциально более обширной и амбициозной перспективы, претендующей на объединение сразу нескольких масштабных теорий функционирования мозга и познания на основе принципов теории самоорганизации и синергетики. Таков предлагаемый Карлом Фристоном «принцип свободной энергии», основывающийся на том известном факте, что биологические самоорганизующиеся системы способны в определенной мере противостоять второму закону термодинамики, сохраняя физиологическую и поведенческую устойчивость (гомеостаз) в условиях непрерывно меняющейся среды. Из этого факта, полагает Фристон, следует, что «… репертуар физиологических и сенсорных состояний, в которых организм может находиться, ограничен, и эти состояния определяют фенотип организма. Математически, –

[1] *Clark A.* Op. cit. P. 185.

пишет Фристон, – это означает, что вероятность данных (интероцептивных и экстероцептивных) сенсорных состояний должна обладать низкой *энтропией*; другими словами, существует высокая степень вероятности, что система будет находиться в любом из небольшого количества состояний, и низкая вероятность, что она будет занимать оставшиеся состояния»[1]. Последние, утверждает Фристон, являются для системы неожиданными и с математической точки зрения относятся к понятию собственной информации (или сюрприза).

Ясно, что подобные состояния в существенной степени зависят от самих агентов – то, что является неожиданным для слона, не будет неожиданным для кита, стрекозы или моржа. Таким образом, совокупность неожиданных состояний в концепции Фристона и составляет для рассматриваемых систем энтропию, которой они вынуждены противостоять. Однако благодаря чему им это удается? Согласно Фристону, ключ к минимизации сюрприза (и его долговременного среднего значения в виде энтропии) для системы лежит в минимизации ее (вариационной) свободной энергии, которая в контексте настоящего рассмотрения обладает тем же самым значением, что и ошибка в предсказании в указанном выше смысле, т.е. как разница между генерируемыми моделью мозга вероятностными предсказывающими репрезентациями мира и сигналами из самого мира, сообщающими о действительном положении дел[2]. Соответственно, «любая самоорганизующаяся система, которая находится в состоянии равновесия со своей средой, должна минимизировать свою свободную энергию»[3].

Каким образом биологические самоорганизующиеся системы могут минимизировать (обобщенно говоря) свои ошибки в предсказании (свободную энергию), чтобы поддерживать устойчивость в изменяющемся мире? Ответ Фристона и сторонников его концепции таков: существует два пути к устранению противоречий между моделью мира и самим миром, а именно (1) либо система

[1] *Friston K.* Op. cit. P. 127.

[2] *Clark A.* Op. cit. P. 186. Во избежание путаницы здесь требуется пояснить, что в концепции Фристона речь идет о так называемой вариационной свободной энергии, понятии (сам термин был введен в лекциях по статистической механике Р. Фейнманом), развитом в области компьютерной науки и машинного обучения в уже упомянутых работах Джеффри Хинтона и других авторов для оценки качества аппроксимации сложного (апостериорного) распределения вероятности P посредством более простого распределения вероятности Q и, таким образом, не имеющем прямого отношения к понятию свободной энергии (свободной энергии Гельмгольца) из классической термодинамики.

[3] *Friston K.* Op. cit. P. 127.

при получении ошибки в предсказании вносит соответствующие коррективы в свою модель мира, обновляя ее с целью порождения более точных предсказаний сенсорного входа, (2) либо система посредством действий и двигательной активности меняет непосредственно сам сенсорных вход, чтобы он в лучшей степени соответствовал генерируемым моделью предсказаниям. Последний способ открывает возможность для байесовской теории (перцептивного) действия, преимущественная цель которого с этих позиций заключается в минимизации той самой пресловутой ошибки в предсказании.

Фристон, таким образом, полагает, что предлагаемая им концепция благодаря ее ключевому императиву минимизации свободной энергии позволяет найти точки соприкосновения и проложить дорогу к (по крайней мере, частичному) объединению сразу нескольких известных перспектив когнитивной науки и нейронауки, а именно теории клеточных ансамблей Д. Хебба, теории корреляций К. фон дер Мальсбурга, теории нейродарвинизма Дж. Эдельмана, теории зрительного внимания как направленного соревнования[1] и др.

Так, например, знаменитый постулат теории Хебба, гласящий, что связи между нейронами усиливаются, если эти нейроны разряжаются совместно, в данном контексте может означать, что в случае строгой корреляции между предсказанием сенсорного входа и сигналом об ошибке и последующего выучивания причинно-следственных отношений, связи между нейронами, кодирующими эти отношения (для генеративной модели), усиливаются. С теорией Эдельмана перспектива минимизации свободной энергии, утверждает Фристон, может быть связана через понятие значения, которое в рамках модели нейродарвинизма использовалось для обозначения врожденных диспозиций и тенденций поведения, необходимых для выживания и адаптивного успеха. Значение (или значимость, англ. value – адаптивно значимое поведение и состояния), указывает Фристон, с позиций развиваемого им подхода может быть понято как обратно пропорциональное сюрпризу (неожиданным состояниям) и кодироваться генеративной моделью с ее априорными ожиданиями, которые и определяют «значение сенсорных состояний и, что важнее всего, являются наследуемыми через генетические и эпигенетические механизмы. Это означает, что априорные ожидания… могут предписать [для агента – М.С.] небольшое число желательных состояний с врожденной значимостью. В свою очередь это позволяет естественному отбору оптими-

[1] *Desimone R., Duncan J.* Neural Mechanisms of Selective Visual Attention // Annual Review of Neuroscience. 1995. Vol. 18. No. 1. P. 193–222.

зировать априорные ожидания и гарантировать, что они согласуются с фенотипом агента»,[1] – пишет Фристон. Наконец, необходимость различения между действительной ошибкой в предсказании и, к примеру, всего лишь шумом или неопределенным сигналом говорит о роли внимания в рамках схемы минимизации ошибок в предсказании сенсорных сигналов. Внимание, поэтому, полагает Фристон и его последователи[2], может быть связано с определением (оптимизацией) точности сигнала, содержащего ошибку в предсказании, так, что сигналы с большей точностью (и, как следствие, надежностью) наделяются большим весом и влиянием на вышестоящие слои вычислительной иерархии.

Теория Фристона, безусловно, заслуживает отдельного основательного обсуждения, однако на этом мы заканчиваем первую часть нашего исследования и переходим к рассмотрению эпистемологической составляющей представленной теории и прежде всего, конечно, к ответу на вопрос, вынесенный в заглавие данной статьи, а именно к вопросу о том, в каком отношении находится байесовская теория восприятия к реалистическому пониманию перцептивных процессов?

Вопрос о реализме и отношение байесовской теории к «ситуативному и воплощенному познанию»: мир неопределенности vs. мир как его собственная лучшая модель

Вопрос о реализме – один из важнейших (если не важнейший) среди вопросов, которые могут встать, когда речь заходит об эпистемологической оценке байесовской теории восприятия. Как уже отмечалось выше, современные байесовские идеи и ТПО являются наследниками теорий Гельмгольца, Грегори, Рока и др., т.е. традиции в изучении восприятия, известной прежде всего как конструктивизм. Таков, к примеру, известный и цитируемый представителями современной байесовской теории фрагмент из работы самого Гельмгольца: «Психические активности, которые ведут нас к заключению, что перед нами в определенном месте существует определенный объект определенного характера являются в общем не сознательными активностями, а бессознательными. По их эффектам они аналогичны *умозаключениям* (*conclusion* – курсив автора – М.С.) в той мере, что наблюдаемое действие на наши чувства позволяет нам формировать идею возможной причины этого действия; хотя на самом деле просто неизменно *нервные стимуляции*

[1] *Friston K.* Op. cit. P. 133.

[2] Подробнее о байесовской теории внимания – см.: *Hohwy J.* Op. cit. 191–206.

воспринимаются прямо, т.е. действия, но никогда не сами внешние объекты (курсив мой – М.С.)»[1].

С другой стороны, наиболее известным реалистическим подходом к восприятию является экологическая теория Дж. Дж. Гибсона[2]. Гибсон, как прекрасно известно, много лет подряд в своих работах отстаивал точку зрения (в философии известную как наивный реализм), что восприятие окружающего мира (и самовосприятие) является прямым, а не опосредованным психическими или физиологическими структурами (изображениями) и процессами (обработкой данных) феноменом. Всем традиционным и современным ему концепциям, отталкивающимся от идеи обработки данных, Гибсон противопоставлял свою так называемую теорию извлечения информации (англ. information pickup). При этом часто забывается, что Гибсон также утверждал, что восприятие может быть и непрямым, например, когда речь заходит о восприятии таких артефактов, как картины (которое, по Гибсону, состоит из прямого восприятия поверхности картины и непрямого осознания того, что на ней запечатлено[3]). Приведем цитату, весьма точно характеризующую взгляды позднего Гибсона: «Что такое прямое восприятие? Когда мы смотрим, скажем, на Ниагарский водопад, а не на картину, на которой он изображен, наше восприятие будет прямым, а не *опосредованным* (здесь и далее курсив автора – М.С.). Опосредованным оно будет во втором случае, когда мы смотрим на картину. Таким образом, когда я утверждаю, что восприятие окружающего мира является прямым, я имею в виду, что оно не опосредованно никаким изображением – *ни сетчаточным, ни нервным, ни психическим*. *Прямое восприятие* – это особый вид активности, направленный на получение информации из объемлющего светового строя. Этот процесс я назвал *извлечением информации*»[4].

Таким образом, теорию Гибсона в известном смысле можно считать едва ли не главным антиподом вдохновленной Гельмгольцем акцентирующей роль знания и априорных ожиданий в восприятии ТПО и байесовской перцептивной психологии. Однако что из этого следует? Оказываются ли ТПО и байесовская перцептивная

[1] *Helmholtz H. von.* Op. cit. P. 4.

[2] *Гибсон Дж.* Экологический подход к зрительному восприятию. М., 1988. См. также: *Gibson J. J.* New Reasons for Realism // Synthese. 1967. Vol. 17. No. 2. P. 162–172. В этой связи особенно интересно прояснить отношение байесовских моделей к испытавшему существенное влияние Гибсона «ситуативному/воплощенному познанию».

[3] *Гибсон Дж.* Указ. соч. С. 408.

[4] *Гибсон Дж.* Указ. соч. С. 213.

парадигма противостоящими реализму в широком смысле? В конце концов, является ли восприятие, как часто отмечается сторонниками байесовской теории (в свете той же концепции функциональной инверсии прямой и обратной сенсорной связи в мозге), своего рода «фантазией», продуктом генеративной модели мозга (и в каком смысле) или даже «контроллируемой галлюцинацией»[1]?

Чтобы попытаться ответить на эти вопросы, обратимся для начала непосредственно к первоисточникам, т.е. к тому, как данная проблема трактуется самими приверженцами ТПО. Так, философ и когнитивист Якоб Хохви, развивающий и конкретизирующий идеи Фристона в своей книге «Предсказывающий разум»[2], в понимании данного вопроса, в общем и целом, следует в русле приведенных выше взглядов Гельмгольца. Восприятие, утверждает он, с точки зрения программы минимизации ошибки в предсказании является непрямым и должно отстоять от мира на расстояние одного шага. Скрытыми за завесой сенсорного входа, указывает он в полном согласии с идеями Гельмгольца, оказываются причины воздействия на наши органы чувств, которые мозг пытается установить в процессе своих каузальных выводов. Порождаемые генеративной моделью богатые репрезентации сенсорного входа, сообщаемые через нисходящие соединения в мозге, с этой точки зрения полностью определяют содержание перцептивного опыта, подавляя предсказанные сигналы, и, как уже неоднократно говорилось, только отличающиеся от предсказанных сигналы могут передаваться наверх с целью внесения корректив в модель мира и, соответственно, порождения более точных предсказаний.

Как указывает Хохви, такого рода непрямой характер восприятия вовсе не означает, что в мозге может существовать пресловутый гомункулюс, созерцающий внутренние репрезентации на некоем ментальном экране. Разумеется, ничего подобного в мозге нет. Это всего лишь означает, говорит он, что то, что мы воспринимаем, определено нисходящими предсказаниями «текущего сенсорного входа, нежели восходящими сигналами положения дел самих по себе»[3]. Более того, с точки зрения Якоба Хохви, именно подобное понимание природы перцептивных процессов способно дать элегантное решение проблеме надлежащих ограничений и контроля (англ. supervision) процесса восприятия. Последнее кон-

[1] *Clark A.* Expecting the World: Perception, Prediction, and the Origins of Human Knowledge [Электронный ресурс]. URL: http://www.research.ed.ac.uk/portal/files/9873993/Expecting_the_World_ACfinalNov2012.pdf (дата обращения: 13.05.2016).

[2] *Hohwy J.* Op. cit.

[3] Ibid. P. 48.

тролируется не посредством некоего могущественного надзирателя (по аналогии с машинным обучением, где используются методы так называемого контролируемого обучения – англ. supervised learning), а при помощи тех самых ошибок в предсказании, которые сообщают мозгу о действительном положении дел, совершенствуя его модель мира. То есть процесс восприятия в данном понимании ограничивается и контролируется непосредственно самим миром, который, как замечает Хохви, собственно, и есть истина[1] (хотя и, как он полагает, доступная нам всегда только через завесу сенсорного входа).

Так и Энди Кларк, соглашаясь с идеей, что содержание нашего перцептивного опыта может определяться нисходящими предсказаниями генеративной модели мозга, тем не менее, отмечает, что все же может быть верно утверждение, «что то, что мы воспринимаем, есть не некая внутренняя репрезентация, а (в точности) мир»[2]. Кларк утверждает, что только при помощи такого характера восприятия может быть установлена тесная связь между мозгом и миром, чтобы наш мозг «отражал и регистрировал релевантные аспекты каузальной структуры самого мира»[3]. На этом основании Кларк обращается к обсуждению отношения байесовской теории к «ситуативному и воплощенному познанию», стремясь продемонстрировать, что эти два направления, несмотря на их кажущуюся противоположность, могут иметь определенные глубинные точки пересечения. Основная идея Кларка заключается в том, что создание и эксплуатирование нами специальных интеллектуальных культурных сред в их самых разных обличиях и формах (в виде тех же слов, изображений, дорожных знаков и т.д.; т.е. того, что так интенсивно исследовалось в рамках модели «ситуативного и воплощенного познания», в том числе и работах самого Кларка) позволяет делать мир более «удобным» для нашего мозга, устанавливая между ними большее соответствие (так сказать, увеличивая их совместную информацию), что в свою очередь позволяет нам более эффективно минимизировать ошибки в предсказании в самых различных контекстах от повседневной жизни до специализированных видов деятельности вроде науки.

Таким образом, мы можем видеть, что, несмотря на то, что, в противоположность Гибсону и прямому реализму в общем, приверженцы программы ТПО утверждают, что восприятие опосредуется (и, более того, в содержательном даже плане определяется)

[1] Ibid. P. 50.

[2] *Clark A.* Whatever next? Predictive brains, situated agents, and the future of cognitive science // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 199.

[3] Ibid.

моделью мира, современная байесовская теория не является противостоящей реализму в широком смысле слова, поскольку, согласно одному из ключевых положений теории, генеративная модель мира постоянно совершенствуется и ограничивается передаваемыми ей по восходящим соединениям сигналами из самого мира, сообщающими о неучтенных ее предсказаниями аспектах среды.

Мы же, в отличие от Гибсона, в целом, симпатизируем идее, что восприятие типичным образом заключает в себе использование ресурсов памяти (в умеренном виде этот тезис не противоречит утверждению, что мы воспринимаем не модели и репрезентации, а сам мир). Основной вопрос здесь, мы полагаем, заключается в той мере, в которой процессы восприятия в действительности полагаются на использование ресурсов и структур знания, а также в том, какого типа структуры оказываются вовлечены в эти процессы. В первом вопросе, как представляется, может крыться главный источник возможных противоречий между ТПО и ставшей уже классической программой «ситуативного и воплощенного познания».

Так, например, достаточно вспомнить, что одной из главных идей всего «ситуативного и воплощенного познания» была идея, удачно выраженная робототехником Родни Бруксом в его знаменитом тезисе: «мир является его собственной лучшей моделью»[1]. Иначе говоря, как указывали Брукс и другие основоположники «ситуативного и воплощенного познания», критикуя классический вычислительный когнитивизм, организму незачем создавать сложные и вычислительно затратные репрезентации и модели мира, если у него есть возможность обратиться (прежде всего, через действия и двигательную активность, включая, конечно, движения глаз и головы) непосредственно к самому миру. Проект «интерактивного зрения» П.С. Черчленд, В. Рамачандрана и Т. Сейновски с их идеей «зрительных полумиров»[2], сходная идея так называемых «частичных репрезентаций» самого Энди Кларка[3], а также более

[1] *Brooks R.* Cambrian Intelligence: The Early History of the New AI. Cambridge, MA, 1999. P. 81, 89, 115, 121, 128, 166–167, 176.

[2] *Churchland P. S., Ramachandran V. S., Sejnowski T. J.* A Critique of Pure Vision // Large-Scale Neuronal Theories of the Brain / Eds. C. Koch, J.L. Davis. Cambridge, MA, 1994. P. 23–60.

[3] *Clark A.* Being There: Putting Brain, Body and World Together Again. Cambridge, MA, 1998. P. 30, 130. См. также: *Сущин М.А.* Концепция ситуативного познания в когнитивной науке: критический анализ: дис. … канд. филос. наук: 09.00.01 / Сущин Михаил Александрович: М., 2014.

поздняя программа «энактивизма»[1] все в равной мере исходили из принципа экономии имеющих место в голове процессов и ресурсов при условии надежной доступности необходимых агентам аспектов внешнего мира.

Как бы сказал сторонник «ситуативного/воплощенного познания», когда я, например, обследую определенную зрительную сцену (даже ту, которую я видел тысячи раз), я не обращаюсь к неким внутренним детальным ее копиям и моделям (которых, как показывают современные интересные эксперименты по «слепоте к изменениям»[2] и «слепоте по невниманию»[3], у меня, собственно говоря, и нет), а просто полагаюсь на быстрые движения своих глаз (как и движения головы и моего тела в общем), направляющих расположенную на сетчатке центральную ямку на обследование тех аспектов доступной мне сцены, которые в данный момент представляют для меня наибольший интерес.

С другой стороны, как постулируется сторонниками ТПО и байесовской парадигмы (в ее более современной, исходящей от Фристона, Кларка и др. версии), при восприятии мозг на самых разных уровнях и шкалах полагается на вероятностные предсказания, производимые генеративной моделью мира, которые, повторимся, согласно этой концепции, определяют содержание нашего перцептивного опыта и подавляют все поступающие на вход из внешнего мира сигналы, кроме тех, которые отличаются предсказанных и которым, соответственно, дозволяется направляться до высших ступенек вычислительной иерархии с целью внесения исправлений в саму генеративную модель. Например, когда мы видим сидящего на коврике кота «Вместо того, чтобы просто репрезентировать «КОТА НА КОВРИКЕ» вероятностный Байесовский мозг будет кодировать условную функцию плотности вероятности, отражающую относительную вероятность данного положения дел (и любых других в определенной степени поддержанных альтернатив) при условии доступной информации…. Сперва система может избегать принятия любой единичной интерпретации, сталкиваясь с начальным потоком ошибочных сигналов… в то время как соревнующиеся «убеждения» распространяются по системе вверх и вниз. После чего обычно следует быстрое согласие на домини-

[1] *Noe A.* Action in Perception. Cambridge, MA, 2004. *Noe A.* Out of Our Heads: Why You Are Not Your Brain and Other Lessons from the Biology of Consciousness. New York, 2010.

[2] *Simons D. J., Resnik R. A.* Change blindness: Past, present, and future // Trends in Cognitive Sciences. 2005. Vol. 9. No. 1. P. 16–20.

[3] *Simons D. J., Chabris C. F.* Gorillas in our midst: sustained inattentional blindness for dynamic events // Perception. 1999. Vol. 28. No. 9. P. 1059–1074.

рующей теме (КОТ, КОВРИК) с дальнейшими впоследствии урегулированными деталями (РАЗНОЦВЕТНЫЙ КОВРИК, ПОЛОСАТЫЙ КОТ). Данная установка, таким образом, предпочитает вид периодически улаживаемых моделей «сути-с-беглого взгляда», где мы вначале определяем общую сцену… с последующими деталями»[1].

Здесь возникает множество вопросов. Наиважнейший, пожалуй, состоит в том, как в рамках единой когнитивной архитектуры мог бы быть распределен баланс сил между описанной выше экономной экологически ориентированной «ситуативной стратегией» и основанной на использовании выводов, моделей, гипотез и знаний (можно сказать, рационалистской) предсказывающей формой когнитивной организации. И хотя может быть верно, что если система в состоянии избегать чрезмерных вычислительных затрат и полагаться на умелое использование ресурсов среды, «она, вероятнее всего, будет поступать подобным образом»[2], общее соотношение двух стратегий, как резонно отмечает Кларк, остается одной из важнейших нерешенных проблем когнитивной науки.

Наконец, между указанными направлениями существует и иное фундаментальное противоречие, которое относится уже непосредственно к пониманию самого мира, который мы воспринимаем. Дело в том, что одно из наиболее базовых допущений байесовской теории восприятия заключается в том, что мир, который воспринимают люди и другие агенты, является миром неопределенности и шума: «Хотя интроспекция говорит нам, что восприятие является детерминистическим и определенным, много факторов вносит вклад в ограничение надежности сенсорной информации о мире – отображение трехмерных объектов в двухмерное изображение, нейронный шум на ранних стадиях сенсорного кодирования и структурные ограничения на нейронные репрезентации и вычисления (например, плотность рецепторов на сетчатке)»[3]. Чтобы нивелировать эффекты неопределенности, противоречивой или плохо сочетаемой сенсорной информации, например, когда две сенсорные модальности (допустим, зрение и слух) предоставляют противоречивую информацию об объекте, по замыслу приверженцев данного направления, и необходим байесовский механизм, который позволяет выработать наиболее вероятную репрезентацию

[1] *Clark A.* Whatever next? Predictive brains, situated agents, and the future of cognitive science // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 188.

[2] *Clark A.* Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 256.

[3] *Knill D. C., Pouget A.* The Bayesian brain: the role of uncertainty in neural coding and computation // Trends in Neuroscience. 2004. Vol. 27. No. 12. P. 712.

воспринимаемого мира с целью эффективного руководства действиями и поведением.

С другой стороны, понимание перцептивного мира в рамках доктрины «ситуативного/воплощенного познания», как уже говорилось, в значительной степени определялось выраженным Бруксом принципом «мира как его лучшей модели». Мир, с этой точки зрения, в большинстве ситуаций является доступным и надежным, мир, как говорит американский философ и один из ведущих теоретиков «энактивизма» Алва Ноэ, «проявляется для нас»[1], т.е. становится открытым в его богатстве и деталях в наших действиях и активном взаимодействии с ним, а не творится в нашем сознании или мозге. Как можно урегулировать данную дилемму?

Мы полагаем, что внести ясность в этот вопрос может различие, идущее от Деннета (хотя некоторые считают, что в известном смысле оно идет еще от Аристотеля) и используемое теоретиками, как в современных исследованиях сознания, так и в когнитивной науке, а именно различие между личностным и субличностным уровнями объяснения. Первый относится к описанию когнитивной системы как целого, которая видит (мыслит, сознает, помнит и т.д.) что-либо. Второй – к работе ее составных частей на всех уровнях (от уровня молекул и нейронов до больших нейронных ансамблей, регионов мозга, глаз и даже, возможно, мозга и нервной системы как целого), кроме наивысшего, т.е. уровня системы как целого. С этой точки зрения, шум и неопределенность, присутствующие для элементов и механизмов субличностного уровня (например, нейронов сетчатки и ранних стадий зрительной обработки), тем не менее, могут не препятствовать выработке вполне определенных и субъективно целостных восприятий, если рассматривать их с позиции личностного уровня описания.

Обильно присутствует шум и неопределенность, вне всяких сомнений, и для когнитивных систем на личностном уровне. Элементарный пример: видимая форма карандаша, опущенного в стакан с водой, искажается (как и видимая форма объектов при некоторых условиях в общем). При просмотре видео на компьютере (в кинотеатре или же просто включив телевизор) мы будем вынуждены соотносить звуковые сигналы, исходящие от динамика, с доступным видеорядом на экране (например, видимыми движениями губ телеведущего), и эти сигналы часто бывают не синхронными. К тому же, как указывал Дж. Дж. Гибсон при обсуждении своей концепции «возможностей», часто «вещи могут выглядеть не та-

[1] *Noe A.* Out of Our Heads: Why You Are Not Your Brain and Other Lessons from the Biology of Consciousness. New York, 2010. P. 3.

ками, каковы они есть»[1]. Так, сулящий проход сквозной проем может оказаться стеклянной дверью, обыкновенный песок – зыбучим, ядовитое растение можно принять за неядовитое и т.д. Для преодоления подобной неопределенности, как уже отмечалось, когнитивные агенты могут полагаться на механизм оптимальной оценки сенсорного входа, наподобие байесовского. И, тем не менее, мы берем на себя смелость утверждать, что в большинстве ситуаций, если подразумевается личностный уровень описания, мир, с которым мы взаимодействуем, отнюдь не является «миром неопределенности», что подтверждается возможностью эффектного действия успешно приспособленных к своим средам агентов и систем. В этом отношении мы можем поддержать сторонников «энактивизма».

Вместо заключения: основные проблемы байесовской программы в исследованиях восприятия и познания

Таким образом, в рамках настоящего обзора мы постарались собрать воедино основные положения и свидительства в пользу нового активно набирающего популярность масштабного направления в современных исследованиях познания – так называемого «предсказывающего кодирования», а также рассмотреть некоторые важные эпистемологические следствия, касающиеся его отношения к реалистическим концепциям восприятия. По словам двух сторонников новой теории, «Байесовские понятия трансформируют исследования восприятия при помощи строгой математической рамки для представления физических и статистических свойств среды, описания задач, которые перцептивные системы пытаются выполнять, и выведения подходящих вычислительных теорий того, как выполнять эти задачи»[2]. Тем не менее, несмотря на всю возможную продуктивность байесовских моделей восприятия, новая программа исследований неизбежно сталкивается с рядом важных вопросов и проблем (как концептуальных и методологических, так и более эмпирического характера), от успешного решения которых в значительной степени и зависят ее дальнейшие перспективы. Часть этих проблем была обозначена выше, а к важнейшим из оставшихся мы планируем перейти прямо сейчас.

[1] *Гибсон Дж.* Указ. соч. С. 211.

[2] Цит. по: *Rescorla M.* Bayesian Perceptual Psychology [Электронный ресурс]. URL: http://www.philosophy.ucsb.edu/docs/faculty/papers/bayesian.pdf (дата обращения: 07.04.2016).

(1) *Происхождение априорных ожиданий*. Часто данный вопрос рассматривается как едва ли не основная проблема байесовской теории, поскольку априорные ожидания в ее рамках играют наиважнейшую роль в селекции гипотез и отсеве заведомо неправдоподобных вариантов. Тривиальный ответ заключается в том, что если подобного рода структуры действительно имеют место быть (что является весьма вероятным), то они, очевидно, должны являться продуктом совместного действия генов и онтогенеза/индивидуального опыта. Но в какой степени? Так, например, известно, что одним из наиболее базовых инструментов восприятия глубины является оценка формы объектов на основе затенения. При восприятии формы на основе затенения мозг, как правило, исходит из допущения (или, в байесовских терминах, априорного ожидания), что источник освещения должен располагаться над головой (англ. the «light-from-above» prior). Является ли оно жестко встроенным или модифицируемым в опыте? Как было показано в работе с говорящим названием «Опыт может изменить априорное ожидание источника света сверху»[1], наше имплицитное допущение направления источника света для восприятия формы на основе затенения является адаптивным и может быть модифицировано в результате обучения в экспериментах, что предположительно свидетельствует, что перцептивные априорные ожидания могут быть постоянно обновляемыми в активном взаимодействии со средой. Тем не менее, относительный вклад среды и генов, как было сказано, остается крайне мало изученным, и в настоящее время исследователи, работающие в рамках байесовской парадигмы, по большей мере пытаются установить состав возможных априорных ожиданий, фактически не затрагивая вопрос об их происхождении[2]. Хотя бы частичное решение этих вопросов, несомненно, могло бы служить существенным подспорьем для развития байесовской теории восприятия.

(2) *Феноменология и границы объяснения*. Из нашего рассказа должно быть ясно, что постулируемые сторонниками байесовской теории предполагаемые механизмы работы процессов восприятия являются, в приведенной выше терминологии Д. Деннета, субличностными и неосознаваемыми, а значит, как представляется, рассматриваемая нами теория не имеет прямого отношения к вопросу о специфике феноменальных, или, как говорят, субъективных, ас-

[1] *Adams W. J., Graf E. J., Ernst M. O.* Experience can change the “light-from-above” prior // Nature neuroscience. 2004. Vol. 7. No. 10. P. 1057–1058.

[2] *Rescorla M.* Bayesian Perceptual Psychology [Электронный ресурс]. URL: http://www.philosophy.ucsb.edu/docs/faculty/papers/bayesian.pdf (дата обращения: 07.04.2016).

пектов восприятия. Тем не менее, для любой фундаментальной теории восприятия вопрос о происхождении феноменальных аспектов перцептивных процессов является одним из тех вопросов, которые не могут просто так быть проигнорированы или оставлены на обочине. Естественный вопрос, поэтому, состоит в том, может ли байесовская теория, собственно говоря, внести какой-либо вклад в понимание природы самого перцептивного опыта?

Наиболее вероятный ответ заключается в том, что байесовская теория, безусловно, может внести вклад в понимание природы перцептивного опыта, но по большей части через описание и объяснение принципов работы нейровычислительных субличностных механизмов, которые ведут к его возникновению. Мы видели, что одно из ключевых положений ТПО гласит, что содержание текущего перцептивного опыта определяется гипотезой, которая «в лучшей степени предсказывает сенсорный вход и получает наибольшую апостериорную вероятность»[1]. Мы видели, что вероятная функциональная роль внимания, как части перцептивных процессов (а значит, и перцептивного опыта), с точки зрения Фристона и его последователей, может заключаться в оценке и селекции сигналов об ошибке в предсказании, где подлинным сигналам, в отличие от просто шума, присваивается наибольший вес и дозволяется подниматься до высших «этажей» вычислительной иерархии с целью внесения корректив в генеративную модель мира.

Подобное понимание функций восприятия и внимания недавно было использовано[2] для рассмотрения с позиций байесовского подхода источников психических расстройств (в особенности, шизофрении). С этой точки зрения, так называемые «позитивные симптомы» шизофрении (галлюцинации и навязчивые состояния) могут быть связаны с нарушением работы механизмов оценки точности ложных сигналов об ошибке (которые предположительно связаны с действием нейромодулятора дофамин). Этим ложным сигналам присваивается ненадлежащий вес и, соответственно, позволяется распространяться вверх по иерархии и вносить «роковые корректировки» в модель мира агентов. Далее весь процесс развивается лавинообразно: искаженная модель мира будет генерировать неадекватные выводы и предсказания о том, что происходит в мире, усиливая сосредоточение субъекта на несущественных вещах и событиях и приводя к формированию странного опыта, характерного для шизофрении (связи наблюдаемых событий самым необычным образом, мании преследования и т.д.). Таким образом, как подчеркивают авторы этого исследования Пол Флетчер и Крис

[1] *Hohwy J.* Op. cit. 201.
[2] *Fletcher P. C., Frith C. D.* Op. cit.

Фрит, язык вычислительной науки в форме байесовской теории может выступить в роли того необходимого концептуального «моста», который должен заполнить «провал» между физическим и психическим уровнями в деле изучения столь нуждающихся в объяснении феноменов человеческого опыта. Если резюмировать сказанное, байесовская теория в большей степени нацелена на прояснение причин возникновения перцептивного (и иных связанных с ним типов) опыта, а не его феноменальной специфики.

(3) *Аргумент Дж. Серла против вычислительных теорий разума и проблема фальсифицируемости байесовских моделей*. Изначально свой аргумент Дж. Серл задумывал против стандартных вычислительных теорий в когнитивной науке и представил в то время, когда возможность байесовского подхода к восприятию только-только обозначилась на горизонте. В связи с развитием новых подходов и направлений не так давно он переадресовал[1] его теории сознания как интегрированной информации, продвигаемой известными нейроучеными Джулио Тонони и Кристофом Кохом.

Схематично представленная, аргументация Серла заключается в том, что используемые в когнитивной науке понятия информации, вычисления, алгоритма, синтаксиса и т.д. задаются «внешними наблюдателями» и в своем существовании зависят от их интерпретации. В то время как, утверждает он, понятия нейробиологии и естественных наук (такие, как, например, «масса», «энергия», «заряд», «фотосинтез» и др.) обозначают свойства реально существующих внутренне присущих миру феноменов и процессов. Сознание же и когнитивные процессы, указывает Серл, не являются зависящими от внешних наблюдателей (я являюсь сознающим вне зависимости от мыслей других людей), и даже более того, их информационная интерпретация оказывается просто бесполезной, когда они, к примеру, получают соответствующее объяснение с позиций нейронауки. Таким образом, отмечает он, остается открытым вопрос, «куда же здесь должны втиснуться формальные символьные манипуляции?»[2]

Как указывали многие авторы, аргументация Серла отнюдь не является безупречной, однако она, тем не менее, затрагивает важнейшую в контексте настоящего рассмотрения тему, а именно что представляют собой объяснения теории, которая постулирует, что восприятие и другие связанные с ним типы опыта являются результатом действия субличностного нейровычислительного меха-

[1] *Searle J.* Can Information Theory Explain Consciousness? [Электронный ресурс]. URL: http://www.nybooks.com/articles/archives/2013/jan/10/can-information-theory-explain-consciousness/ (дата обращения: 16.09.2015).

[2] *Серл Дж.* Открывая сознание заново. М., 2002. С. 186.

низма, чья цель заключается в поддержании равновесия организма с его средой посредством минимизации ошибок в предсказании сенсорных сигналов? Каков научный статус теории, пытающейся раскрыть принципы функционирования этой машины в терминах математической теории вероятности и вариационного байесовского исчисления? Является ли байесовская теория просто удобным инструментом исследования работы субличностных механизмов восприятия? Является ли байесовская теория адекватным инструментом исследования работы механизмов восприятия?

Эти и другие вопросы недавно получили освещение в нескольких объемных критических обзорах байесовской программы в когнитивной науке[1], где утверждалось, что типичные байесовские модели восприятия и когнитивных процессов в силу их чрезмерной общности и способности подстроиться практически под любые данные являются описательными (а не объяснительными) и нефальсифицируемыми. Помимо этого, были поставлены под сомнение ключевые для байесовской теории положения о чрезмерном шуме/неопределенности в процессе сенсорного кодирования сигналов (что в случае отсутствия такового делает байесовский механизм попросту ненужным), а также общем оптимальном (т.е. более или менее приближенно следующем рациональным стандартам байесовских вычислений) характере человеческого познания (что, как было замечено, является весьма сомнительным в свете известных данных о далеком от совершенного или сколько-нибудь оптимального характере эволюции когнитивных агентов).

Как бы то ни было, мы придерживаемся точки зрения, что развитие когнитивной науки невозможно только на пути аккумуляции эмпирических (поведенческих и нейробиологических) данных, но нуждается в адекватной теории на уровне общей цели и содержания деятельности когнитивных агентов и систем, без которой эмпирический массив данных будет лишь скоплением фактов. Байесовская теория на текущий момент является наиболее целостной и систематичной современной программой исследований в этой области, хотя и, безусловно, нуждающейся в дальнейшей разработке и развитии.

(4) *«Заземление» в эмпирических исследованиях*. То, что надлежащее развитие исследований когнитивных процессов невозможно без развития теории собственно когнитивных процессов, было известно, разумеется, еще и до возникновения современной байесовской теории восприятия. Как постулировал в свое время известный ученый в области изучения нейровычислительных основ зрения

[1] См., например: *Bowers J. S., Davis C. J.* Bayesian Just-So Stories in Psychology and Neuroscience // Psychological Bulletin. 2012. Vol. 138. No. 3. P. 389–414.

Дэвид Марр[1], любое надлежащее исследование зрительных процессов должно вестись минимум на трех уровнях: уровне вычислительной теории (связанном с определением общей цели вычислительного процесса), уровне алгоритмов (ответственном за нахождение конкретных вычислительных средств реализации теории) и уровне физической реализации теории и алгоритмов в мозге или компьютере. В духе своего времени Марр придерживался точки зрения, что исследования на уровне вычислительной теории, хотя и нуждаются в корреляции с исследованиями физического уровня, тем не менее, являются приоритетными и могут опережать изучение деталей реализации.

Как прекрасно известно, с тех пор ситуация успела измениться кардинальным образом, и теперь уже задача «заземления» моделей в конкретном материале нейронауки стала едва ли не важнейшей для всего поля когнитивных исследований. К сожалению, хотя и количество эмпирических исследований, направленных на поддержку байесовских моделей восприятия, постепенно увеличивается[2], пока что рассмотренная нами теория далека от необходимого для ее полного оправдания объема эмпирических данных. Мы видели, что вероятными претендентами на реализацию цикла порождения высокоуровневых гипотез для предсказания текущего сенсорного входа/вычисления неучтенных параметров среды и прямого сообщения ошибки верхним «этажам» иерархии, согласно предположению Д. Мамфорда, могут являться глубокие и поверхностные пирамидальные клетки регионов коры мозга. В другой рассмотренной нами схеме, предложенной Р. Рао и Д. Баллардом, предсказывающие алгоритмы могут быть вовлечены в активность нейронов ранних зрительных кортикальных областей (к примеру, зоны V1) с так называемыми экстраклассическими эффектами рецептивных полей. Иные иллюстрации включают в себя возможную ведущую роль нейромодулятора дофамин в оценке надежности сигналов об ошибке, динамическое предсказывающее кодирование естественных сцен нейронами сетчатки[3], а также целый ряд других преимущественно непрямых поведенческих и нейрофизиологических свидетельств в пользу действенности байесовских моделей

[1] *Марр Д.* Зрение. Информационный подход к изучению представления и обработки зрительных образов. М., 1987.

[2] *Pouget A., Beck J. M., Ma W. J., Latham P. E.* Probabilistic brains: knows and unknowns // Nature neuroscience. Vol. 16. No. 9. P. 1170–1178.

[3] *Hosoya T., Baccus S. A., Meister M.* Dynamic predictive coding by the retina // Nature. 2005. Vol. 436. No. 7047. P. 71–77.

восприятия и ТПО[1]. Тем не менее, как отмечали многие авторы, выступившие с комментариями на программную статью Э. Кларка в журнале «Науки о поведении и мозге», текущим байесовским построениям недостает множества деталей, касающихся конкретных вычислительных и нейронных механизмов, которые могли бы реализовывать общую схему минимизации ошибки в предсказании. Нахождение этих деталей является главным эмпирическим вызовом для байесовской теории восприятия.

В заключение мы должны сказать, что представленная в рамках настоящего обзора теория является строгим, элегантным, систематическим и многообещающим путем к построению подлинной фундаментальной теории восприятия, совместимой с реалистической эпистемологической платформой в широком смысле. Однако чтобы оправдать выданные ей щедрые авансы байесовская теория восприятия должна прежде всего эффективно ответить на тот немалый круг концептуальных и эмпирических вызовов, часть из которых была приведена в данной статье.

Список литературы

Брунер Дж. О готовности к восприятию // Брунер Дж. Психология познания. М.: Издательство «Прогресс», 1977. С. 13–64.

Гибсон Дж. Экологический подход к зрительному восприятию. М.: Прогресс, 1988. 464 с.

Кант И. Критика чистого разума. М.: Мысль, 1994. 591 с.

Марр Д. Зрение. Информационный подход к изучению представления и обработки зрительных образов. М.: Радио и связь, 1987. 400 с.

Серл Дж. Открывая сознание заново. М.: Идея-Пресс, 2002. 256 с.

Сущин М.А. Концепция ситуативного познания в когнитивной науке: критический анализ: дис. … канд. филос. наук: 09.00.01 / Сущин Михаил Александрович: М., 2014. 137 с.

Adams W. J., Graf E. J., Ernst M. O. Experience can change the “light-from-above” prior // Nature neuroscience. 2004. Vol. 7. No. 10. P. 1057–1058.

[1] Подробнее – см.: *Clark A.* Whatever next? Predictive brains, situated agents, and the future of cognitive science // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 191–192.

Bowers J. S., Davis C. J. Bayesian Just-So Stories in Psychology and Neuroscience // Psychological Bulletin. 2012. Vol. 138. No. 3. P. 389–414.

Brooks R. Cambrian Intelligence: The Early History of the New AI. Cambridge, MA: A Bradford Book/The MIT Press, 1999. 213 pp.

Churchland P. S., Ramachandran V. S., Sejnowski T. J. A Critique of Pure Vision // Large-Scale Neuronal Theories of the Brain / Eds. C. Koch, J.L. Davis. Cambridge, MA: A Bradford Book/The MIT Press, 1994. P. 23–60.

Clark A. Being There: Putting Brain, Body and World Together Again. Cambridge, MA: A Bradford Book/The MIT Press, 1998. 292 pp.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 182–183.

Clark A. Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver // Behavioral and Brain Sciences. 2013. Vol. 36. No. 3. P. 233–244.

Clark A. Expecting the World: Perception, Prediction, and the Origins of Human Knowledge [Электронный ресурс]. URL: http://www.research.ed.ac.uk/portal/files/9873993/Expecting_the_World_ACfinalNov2012.pdf (дата обращения: 13.05.2016).

Dayan P., Hinton G. E., Neal R. M., Zemel R. S. The Helmholtz Machine // Neural Computation. 1995. Vol. 7. No. 5. P. 889–904.

Desimone R., Duncan J. Neural Mechanisms of Selective Visual Attention // Annual Review of Neuroscience. 1995. Vol. 18. No. 1. P. 193–222.

Fletcher P. C., Frith C. D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia // Nature Reviews Neuroscience. 2009. Vol. 10. No. 1. P. 48–58.

Friston K. The free-energy-princple: a unified brain theory? // Nature Review Neuroscience. 2010. Vol. 11. No. 2. P. 127–138.

Friston K. What Is Optimal About Motor Control // Neuron. Vol. 72. No. 3. P. 488–498.

Gibson J. J. New Reasons for Realism // Synthese. 1967. Vol. 17. No. 2. P. 162–172.

Gregory R. L. Perceptions as Hypotheses // Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 1980. Vol. 290. No. 1038. P. 181–197.

Gregory R. L. Knowledge in perception and illusion // Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 1997. Vol. 352. No. 1358. P. 1121–1128.

Helmholtz H. von. Concerning the Perceptions in General // Helmholtz H. von. Treatise on physiological optics. Vol. 3. Ch. 26. 3rd edn.

Translated by J.P.C. Southall, 1925, *Op. Soc. Am.* Section 26, reprinted New York: Dover, 1962. P. 1–37.

Hinton G. E. Learning multiple layers of representation // Trends in Cognitive Sciences. 2007. Vol. 11. No. 10. P. 428–434.

Hinton G. E. To Recognize Shapes, First Learn to Generate Images [Электронный ресурс]. URL: http://www.cs.toronto.edu/~hinton/absps/montrealTR.pdf (дата обращения: 05.04.2016).

Hinton G. E., Dayan P., Frey B. J., Neal R. M. The «wake-sleep» algorithm for unsupervised neural networks // Science. 1995. Vol. 268. No. 5214. P. 1158–1161.

Hohwy J. The Predictive Mind. New York: Oxford University Press, 2013. 288 pp.

Hohwy J., Roepstorff A., Friston K. Predictive coding explains binocular rivalry: An epistemological review // Cognition. 2008. Vol. 108. No. 3. P. 687–701.

Hosoya T., Baccus S. A., Meister M. Dynamic predictive coding by the retina // Nature. 2005. Vol. 436. No. 7047. P. 71–77.

Kanizsa G. Seeing and thinking // Acta Psychologica. 1985. Vol. 59. No. 1. P. 23–33.

Knill D. C., Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation // Trends in Neuroscience. 2004. Vol. 27. No. 12. P. 712–719.

Lee T. S., Mumford D. Hierarchical Bayesian inference in the visual cortex // Journal of the Optical Society of America. 2003. Vol. 20. No. 7. P. 1334–1348.

Mumford D. On the computational architecture of the neocortex: I The role of the thalamo-cortical loop // Biological Cybernetics. 1991. Vol. 65. No. 2. P. 135–145.

Mumford D. On the computational architecture of the neocortex: II the role of cortico-cortical loops // Biological Cybernetics. 1992. Vol. 66. No. 3. P. 241–251.

Neisser U. Cognitive Psychology. New York and London: Taylor and Francis, 2014. 348 pp.

Noe A. Action in Perception. Cambridge, MA: The MIT Press, 2004. 296 pp.

Noe A. Out of Our Heads: Why You Are Not Your Brain and Other Lessons from the Biology of Consciousness. New York: Farrar, Strauss and Giroux, 2010. 232 pp.

Palmer S. Vision Science: Photons to Phenomenology. Cambridge, MA: A Bradford Book/The MIT Press, 1999. 832 pp.

Perception as Bayesian Inference / Edited by D. C. Knill, W. Richards. New York: Cambridge University Press, 1996. 532 pp.

Pouget A., Beck J. M., Ma W. J., Latham P. E. Probabilistic brains: knows and unknowns // Nature neuroscience. Vol. 16. No. 9. P. 1170–1178.

Rao R. P. N., Ballard D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects // Nature neuroscience. 1999. Vol. 2. No. 1. P. 79–87.

Rescorla M. Bayesian Perceptual Psychology [Электронный ресурс]. URL: http://www.philosophy.ucsb.edu/docs/faculty/papers/bayesian.pdf (дата обращения: 07.04.2016).

Rescorla M. Bayesian Sensorimotor Psychology [Электронный ресурс]. URL: http://www.philosophy.ucsb.edu/docs/faculty/michael-rescorla/rescorla_bayesian-sensorimotor-psychology.pdf (дата обращения: 09.04.2016).

Rock I. The Logic of Perception. Cambridge, MA: The MIT Press, 1983. 378 pp.

Searle J. Can Information Theory Explain Consciousness? [Электронный ресурс]. URL: http://www.nybooks.com/articles/archives/2013/jan/10/can-information-theory-explain-consciousness/ (дата обращения: 16.09.2015).

Seth A. The Cybernetic Bayesian Brain [Электронный ресурс]. URL: http://open-mind.net/papers/the-cybernetic-bayesian-brain (дата обращения: 21.07.2016).

Simons D. J., Chabris C. F. Gorillas in our midst: sustained inattentional blindness for dynamic events // Perception. 1999. Vol. 28. No. 9. P. 1059–1074.

Simons D. J., Resnik R. A. Change blindness: Past, present, and future // Trends in Cognitive Sciences. 2005. Vol. 9. No. 1. P. 16–20.