Карта и территория

Рациональность: от ИИ до зомби. Том 1
Клуб чтения цепочек в Кочерге

Мы прочитали посты. Давайте выпишем такие штуки:
- что теперь будем делать иначе
- что осталось непонятно
- с чем не согласны
- всё новое и ценное

Ссылка на конспект второго тома:
https://docs.google.com/document/d/1eH5THY75tvMwQ7vBrL5Dx53Of2whPRoLSqKRP7Gq2io/edit?usp=sharing

Ссылка на конспект третьего тома:
https://docs.google.com/document/d/1MIJmLkw0Mng14vzqLDoHT-pJL9h-zyRtVvOtCgziqW8/edit#heading=h.etw49zej3tdn

Ссылка на конспект четвертого тома:
https://docs.google.com/document/d/16gZ92B2B7fuP6VZA6_VWZS1d1g-HTZOtw8LbPETUtFo/edit#

Ссылка на конспект пятого тома:
https://docs.google.com/document/d/1KBnqV6JPtLFtxBKRBhiHZlYNeiN1dS7BRdAWSJiWHjQ/edit#

Ссылка на конспект шестого тома:
https://docs.google.com/document/d/14sFUK7TrD-CxQYNGvaI45AQZ999yWZNK8dgjgJoH3nM/edit#

Ссылка на конспект постов с LessWrong:
https://docs.google.com/document/d/1v-lb3UnMBd2KF-WVGQgkpxCmqciiQbFtfOF-ycvVSDI/edit?usp=sharing

Предисловие

Искажения: введение

Предсказуемо неправы

1. Что такое рациональность

2. Мне сегодня рационально

3. Зачем нужна истина?

4. Что же такое искажение?

5. Доступность

- Эвристика доступности. То, что проще удается вспомнить, кажется более вероятным.
- Связано с тем, что относительно не так давно (когда жили племенами) такой способ оценки окружающего мира был не плох. Информации было мало, и получить ее можно было из первых рук. Огласку получает в основном то, что более необычно и вызовет больший отклик.
- Проблема: Те, кто пережили небольшое наводнение, устанавливают полученный ущерб как верхнюю границу. И будучи защищенными от такого ущерба (например, дамбой), не задумываются об ущербе большего масштаба.

6. Обременительные детали

- Суждение о вероятности события подменяется суждением о его правдоподобности.
- Правдоподобность основывается на том насколько описание события похоже на уже случившиеся события. Прошедшие события мы знаем в деталях и таких же деталей ожидаем от прогнозируемого события.
- Как избежать:
    - замечать перечисление деталей («и»)
    - оценивать вероятность каждой детали отдельно или, по крайней мере, штрафовать за наличие «и», за более длинное описание
    - задаваться вопросом «Откуда мы знаем эту деталь?»
    - обобщать (неопределенный круг причин события вероятнее какой-то одной конкретной причины)

7. Ошибка планирования

- Прогноз при самом удачном стечении обстоятельств и прогноз с учётом всего, что может пойти не так почти не отличаются.
- В то время как реальность преподносит обстоятельства хуже, чем «самое неудачное стечение обстоятельств», какое мы обычно воображаем.
- Как избежать:
    - Использовать взгляд снаружи. Т.е. избегать мыслей о конкретных, уникальных деталях проекта, и просто узнать, сколько времени заняло завершение похожего проекта в прошлом.

8. Иллюзия прозрачности: почему вас не понимают

- Мы знаем, что означают наши слова, и ожидаем, что остальные тоже это знают

9. Ожидая короткие понятийные расстояния

- Основной аргумент: учитывая искажения, уравняйся с аудиторией, затем помоги ей сделать логические умозаключения, чтобы прийти к твоему аргументу
- Почему понятийные расстояния важны?
    - Аудитория не просто невосприимчива к вещам, которые она не понимает - она, скорее, посчитает тебя заносчивым или сумасшедшим
    - Если аргумент на много шагов далек от аудитории, закономерно будет не предполагать, что ты искренне умен и прав - есть риск того, что ты обманщик
- Искажения: иллюзия прозрачности, самоякорение
- Эволюционное объяснение - это “just-so-story” пока не доказано быть проверяемым (хотя иногда такой термин просто служит тому, чтобы обесценить гипотезу)
- Пренебрежение приводит к тому, что образование становится элитарным

10. Линза, видящая свои изъяны

- Мы можем понять видение
- Наука: понятный процесс в мире, который коррелирует мозг с реальностью
- Когда мы впитываем проверяемые ожидания насчет окружающего мира, нужно учесть присущие нам внутренние искажения
- Мозг человека - неточная линза, которая может увидеть свои изъяны - свои систематические ошибки - и добавить коррекции второго порядка, чтобы их влияние уменьшить
- Какие искажения мы не можем заметить в принципе?

    - Если что-то существует вне восприятия всего человечества - изобретена математика или открыта? все равно у нас ничего лучше мозга нет, так что неважно
    - Если что-то другие люди могут заметить
    - Теорема Геделя (карта - это не территория): любая достаточно мощная полная система содержит ограничение - не все математические утверждения можно будет доказать

О цепочке целиком

- Четкое объяснение, что такое рациональность вообще
- Метафора с картой и территорией
- Wimmer and Perner 1983, Sally-Ann or Location False Belief Task: нужно переключиться на знание того, что другие люди не знают того, что мы знаем
- Ариэли

Ложные убеждения

11. Убеждения должны окупаться

Конспект

- Мы можем смоделировать то, что не можем увидеть - необязательно чувственный опыт
- Ошибка предков: флогистон был назван источников огня только после того, как видели огонь
- Эмпирицизм: постоянно спрашивать о том, какой опыт наши убеждения предсказывают - или (даже лучше) запрещают
- Непривязанные убеждения не ограничивают ожидаемый опыт

Обсуждение

- Предсказательная способность и фальсифициерумость
- Мы можем учить физику и решать физические задачи с отличием на олимпиадах, но необязательно понимать связь физики с реальностью

12. Сказ о науке и политике

Конспект (пересказ, инсайты, вот это всё)

- вымышленная история об обществе будущего, живущем в подземельях и никогда не видевшего неба
- спор между Синими и Зелеными насчет цвета неба
- разные исходы столкновениям со свидетельством, что небо синее:

    - Адитья Синяя: начать войну с Зелеными
    - Бэррон Зелёный: вселенная несправедлива
    - Чарльз Синий: спрятать свидетельство
    - Дарья: сменить убеждение
    - Эддин Зелёный: мир полон глупцов
    - Феррис: исследовать новый мир
    - Лоретта Зелёная: отрицание явления
    - Джон Экуменист: притворная мудрость
- описательный пример нормативного принципа "keep your identity small"
- возможные реакции того, как люди неправильно работают с убеждениями, плавающие убеждения
- Юдковский хотел показать, насколько глупы системы убеждений, которые не имеют отношения к свидетельству - исследователю важен мир, а не микросоциум - архетип человека на LessWrong
- групповая идентичность мешает решать проблемы, если цель - находить ответы
- спорный эксперимент Робберс Кейв с двумя лагерями школьников, которые быстро обрели разные групповые идентичности: https://ru.wikipedia.org/wiki/Летний_лагерь_(эксперимент)
- Реальные примеры ситуаций с синими и зелёными и про неважность «цвета неба» в идеологии: https://www.lesswrong.com/posts/xtHd6sfdr2bZHa6Pb/the-ideology-is-not-the-movement

Вопросы

1) Что хотел сказать автор?)) Ваши версии главной мысли/идеи текста. Вариант: какую идею вы извлекли для себя (не претендуя, что Юдковский именно это и хотел приоритетно донести)
2) Какие параллели можно провести с нашей реальностью? Есть ли такие примеры, где чтение этого отрывка/знание из него поможет более объективно и продуктивно решить задачу? Где могли бы быть такие примеры в гипотетических ситуациях? Задачи как глобального масштаба, общечеловеческих ценностей так и свои частные вопросики)

---

13. Вера в убеждения

Конспект (пересказ, инсайты, вот это всё)

- Карл Саган: "плохая гипотеза должна ловко маневрировать, чтобы избежать опровержения"
- Юдковский: плохая гипотеза может быть основана на вере в убеждение
- Дэниел Деннет: "когда трудно верить в X, намного легче верить в то, что ты обязан верить в X"

- Юдковский: рациональность второго уровня невозможна - успешный самообман невозможен (https://lesswrong.ru/w/Двоемыслие_выбирая_быть_искаженным)
    - Соарес писал про намеренный обман своей первой системы как технику продуктивности http://mindingourway.com/dark-arts-of-rationality/
    - статья про кризис веры - как у себя находить и проверять убеждения: lesswrong.ru/423
- По поводу прошлой темы пост - основная мысль: «Религия это не буквально о вере в Бога», «идеология не о том, о чём она номинально в определении определена» https://www.lesswrong.com/posts/xtHd6sfdr2bZHa6Pb/the-ideology-is-not-the-movement
- с религиозными людьми не поспорить о том, что бога нет, но можно поспорить о том, какую мораль они хотят видеть в обществе
- даже жизнью самозатворника можно замечать у себя ложные убеждения - необязательно другие люди

Вопросы

1) Как выглядит правильная модель ситуации, контролирующая ожидания человека, проводящего эксперимент (которую советует Карлу Юдковский)? Сделать гипотезу фальсифицируемой?

---

14. Байесианское дзюдо

Конспект (пересказ, инсайты, вот это всё)

- теорема Ауманна: два рационалиста не могут согласиться не соглашаться
- можно поспорить о чем-то более рациональном - не обязательно спорить о том, есть ли бог или нет

Вопросы

1) С какой целью можно использовать байесианское дзюдо?) Реально на минуту показалось, что в тексте написано: “Женщина, которая стояла рядом и слушала наш разговор, серьёзно посмотрев на меня, сказала: «Это было прекрасно». Немного помявшись, она положила руку мне на ширинку”. Религиозного типа с обеда Юдковский явно не переубедил, да и вообще это победа над грудничком. Окей, тот заблуждается, но как это поможет рационалисту, если у его собеседника нет мотивации работать над убеждениями, а точнее, в каких случаях может пригодится теорема Ауманна? Является ли (косвенно) призыв остаться каждому при своем признаком отсутствия убедительных аргументов? Или, скажем, появление у себя интенции закончить дискуссию именно такой фразой как показатель беспомощности?

Обсуждение

- Теорема Ауманна: два рационалиста не могут согласиться не соглашаться
- Можно посмотреть о чем-то более рациональном - не обязательно спорить о том, есть ли бог или нет
- Юдковский использует приемы, которые помогают собеседнику сформулировать свое убеждение
- Дзюдо - потому что Юдковский заставляет собеседника спорить с самим собой

---

15. Притворная мудрость

Конспект (пересказ, инсайты, вот это всё)

- Паулу Фрейре: "Умывать руки во время конфликта между сильным и бессильным – значит встать на сторону сильного, а не быть нейтральным."
- нейтральность - это точное суждение и также является позицией в обсуждении
- разумно тратить силы на обсуждения, от которых может быть больше пользы, чем на те, в которых множество участников тратит много сил
- Притворная мудрость - боже, как я умен, раз не вмешиваюсь в эти мелочные споры

Обсуждение

- Отказ от выбора не отмажет от последствий того или иного выбора
- Разница между:
    - Принятием нейтрального суждения
    - Отказом вложить незначительные ресурсы
    - Притворной мудростью
- В реальной жизни соблазнительно не выбирать

Ссылки из обсуждения:

- https://ru.wikipedia.org/wiki/Летний_лагерь_(эксперимент)
- https://lesswrong.ru/w/Двоемыслие_выбирая_быть_искаженным
- http://mindingourway.com/dark-arts-of-rationality/
- https://lesswrong.ru/w/Кризис_веры
- https://www.lesswrong.com/posts/xtHd6sfdr2bZHa6Pb/the-ideology-is-not-the-movement
- https://www.mann-ivanov-ferber.ru/assets/files/bookparts-new/strategicheskie-igryi/Strategicheskie_igry_mail_stamped.pdf

---

16. Претензии религии на неопровергаемость

Конспект

- В былые дни люди действительно верили в то, что говорила им их религия, а не просто считали религию важной. Библейские археологи, отправившиеся искать Ноев Ковчег, не считали, что впустую тратят своё время.
- Религия содержала в себе всё: историю, мораль, право, описание мира. Отрыв религии от фактов произошёл позднее, когда оказалось, что факты не стыкуются с религией.
- Современная концепция религии как чего-то строго *морального* порождена тем, что все остальные сферы были отняты у неё более компетентными институтами. Однако и мораль не защищена от человеческого прогресса.
- Идея, что религия это отдельный магистерий, который ни доказать, ни опровергнуть - это то, что повторяют, не задумываясь. Это также и сильное искажение исторического предназначения религии.
- у религии со временем осталась только мораль и ушли другие области (история, наука…)
- аллегорическое толкование Библии о морали не соответствует буквальному толкованию текста
    - для иудеев часть про что есть понимается буквально до сих пор
- мораль в православии: бог считается высшим созданием, который волен поступать с людьми, как хочет
- цель эссе:
    - полемика с точкой зрением о том, что религию нельзя ни подтвердить, ни опровергнуть
    - можно смотреть на религию как на веру, но также события, описанные в ней - рациональны. те, кто писал Библию, имел рациональные ожидания, которые со временем отваливались, потому что они не подтвердились
- в математике это отдельный тип доказательства: нельзя ни доказать, ни опровергнуть

Вопросы

- “Но вы найдёте множество вполне научных заявлений (English), вроде Вселенной, созданной за шесть дней (что является метафорой Большого Взрыва), или кроликов, жующих жвачку (что является метафорой…).” Про кроликов это метафора чего?
    - В библейских установлениях относительно дозволенных для употребления в пищу животных сказано, что нельзя есть *«зайца, потому что он жуёт жвачку, но копыта у него не раздвоены, нечист он для вас»* (Лев. 11:6) - на самом деле, кролики не жуют жвачку.

Обсуждение

- Утверждение о том, что религия — отдельный магистерий, недоступный ни подтверждению, ни опровержению, защищает религию от критики науков

- Независимость религии от фактических реалий — весьма недавняя и исключительно западная концепция.
- Ветхий завет писался во время, когда можно было сказать что угодно, а во время Нового завета уже можно было проверять факты
- У религии со временем осталась только мораль и ушли другие области (история, наука…)
- Аллегорическое толкование Библии о морали не соответствует буквальному толкованию текста
    - Для иудеев часть про что есть понимается буквально до сих пор
- Мораль в православии: бог считается высшим созданием, который волен поступать с людьми, как хочет
- Цель эссе:
    - Полемика с точкой зрением о том, что религию нельзя ни подтвердить, ни опровергнуть
    - Можно смотреть на религию как на веру, но также события, описанные в ней - рациональны. те, кто писал Библию, имел рациональные ожидания, которые со временем отваливались, потому что они не подтвердились
- В математике это отдельный тип доказательства: нельзя ни доказать, ни опровергнуть

---

17. Провозглашения и крики ободрения

Конспект

- Женщина-язычница на конференции рассказывала абсурдную историю про сотворение мира. Явно была довольна собой, но при этом не старалась быть убедительной. Игра на толерантности и противоречии здравому смыслу доставляла ей удовольствие.
    - Мета: Меня смущает последнее предложение. Не стал бы утверждать сходу, что женщина осознанно играет и получает удовольствие (проявления психопатического, манипулятивного поведения).
- Привлечение внимания абсурдностью суждений.

Обсуждение

- Для людей важно не верить во что-то, а демонстрировать свою веру - заявить принадлежность к общности
- Провозглашение - самоопределение, желание выделиться из аудитории (необходима аудитория), привлекает внимание
- Примеры провозглашения:
    - https://www.youtube.com/watch?v=QozfotPOXdY
    - любое критическое утверждение

    - обсуждение в Кочерге на тему "было бы хорошо, если бы профессор Квиррелл правил миром" - люди соглашаются, что пусть он убьет много идиотов - эти люди принимают рациональность, как высшую ценность - это поломанное убеждение
    - эссе Скотта про ярость: https://slatestarcodex.com/2014/12/17/the-toxoplasma-of-rage/
- Это риторический прием: http://esr.ibiblio.org/?p=1068:
    - In Professing and Cheering. you write “That’s why it mattered to her that what she was saying was beyond ridiculous. If she’d tried to make it sound more plausible, it would have been like putting on clothes.” I found it extremely odd that you did not fully understand what you were seeing, but perhaps that is only because I am a neopagan myself and used to pulling similar maneuvers. Or maybe you have borderline Aspergers or something and are poorly equipped to process some kinds of neurotypical interaction, including this one (that’s my wife Cathy’s guess, and not a hostile one; she rather likes you).
    - Your lady panelist was performing a mindfuck. The intent of her speech acts was not to persuade anyone that she believed the Norse creation myth, it was to hold up a funhouse mirror to the religious cognitive style. The question her provocation was implicitly posing to the audience is “If you reject this as absurd, on what basis do you maintain your own equally poetic and absurd creation myth?”
    - I speak from the authority of direct personal experience here, as I have done the same sort of thing for the same reasons in pretty much the same way.

---

18. Убеждение как одеяние

Конспект

- Убеждение как групповая идентификация, способ входить в сообщество. Ты знаешь что нужно говорить и чего говорить нельзя, чтобы быть “своим”.
- Убеждение становится одеянием, когда его суть тебе не важна, но ты знаешь, что “просто так принято” в этом обществе.
- Пример: американцы считают террористов трусами (это одеяние американца). Говорить, что террористы - герои в рамках своей культуры (именно так террористы сами себя видят), значит надевать одеяние врага.
- Идентификация с племенем — очень мощная эмоциональная сила, люди готовы за неё умереть

Обсуждение

- Убеждения: контроллеры ожиданий (полноценные), веру в убеждения, провозглашения и крики ободрения , убеждение как групповая идентификация, способ входить в сообщество (неполноценные)
- Контроллеры ожидания: убеждения, которые сцеплены с реальностью, и реальность может их опровергнуть, определяют действия сейчас
    - Гравитация, но не современное христианство (можно понимать аллегорически, может объяснить, что угодно)
    - Многие христиане ожидают, что после смерти попадут в рай или в ад
- Нужно расцепить контроллер поведения и контроллер ожидания?
    - Есть убеждения, которые не порождают ожидания?
    - Какое бы ни было убеждение, оно порождает ожидание
    - Могут быть просто поломанные убеждения, не привязанные к ожиданиям
- Полноценное убеждение может быть неверным или иррациональным, но остальные формы иногда трудно вообще считать за убеждения.
- Избавиться от всех групповых идентичностей не получится и может быть не очень полезно
    - Сапольски: самоидентификация уменьшает тревожность людей
    - Психика стремится к отсутствию неопределенности и консистентности, сбережению когнитивных ресурсов
    - Групповые идентичности сложились эволюционно
- Посты про сообщество у Скотта: должен быть максимальный ярко выраженный уникальный сигналинг, который не имеет никакой другой нагрузки и не будет проинтерпертирован иначе
    - Например, иудейская кипа, крест, свастика (одеяние), феминитивы - пример убеждения как одежды

---

19. Табличка “Аплодисменты”

Конспект

- Способ получить респект, сказав что-то “правильное” (но не наполненное содержанием).
- В большинстве случаев таблички «Аплодисменты» прямолинейны и могут быть выявлены простым реверсивным тестом - если обратное звучит абсурдно, следовательно, не перевернутое утверждение, возможно, нормально, подразумевая, что это не несет новой информации.
    - «Мы должны сбалансировать риски и возможности ИИ»
    - «Мы не должны соблюдать баланс рисков и возможностей ИИ» - звучит заведомо абсурдно
- Оценка ерундой должна быть основана на ценностях - это не является тривиальным

- Если за утверждением типа «Мы должны сбалансировать риски и возможности ИИ» не следует конкретики (планы, действия, критика), то это именно табличка.
- Грубо говоря, Юдковский критикует наполнение текстов “водой”.

Обсуждение

- Какие цели человек преследует? - Политический популизм
- ТА могут быть полезны не для передачи информации
- Может ли быть смысл в том, что не сказано, а ТА используется для отвлечения внимания
- Скотт: переворачивайте любой совет, который слышите
- ТА может навредить аудитории и говорящему

Замечая замешательство

20. Сфокусируй неуверенность

Конспект

- легче объяснять события, которые уже произошли, чем те, что будут
- применение математики и крит. мышления вне стен университета
- некоторые исходы мы ожидаем больше, чем другие
- априорные представления
- Юдковский вводит концепт вероятности: рационально распределять свои ресурсы в зависимости от вероятностей исходов
    - байесианское понимание вероятности: мера нашего незнания
    - любая вероятность: ожидаемая
    - вероятность/неопределенность: свойство карты
- если есть неопределенность - оценить, какие исходы мы ожидаем больше или меньше
    - начать с lower and upper bounds (какие-то данные всегда есть)
- фундаментально невозможно достать больше ожидания
- следует:
    - выражать ожидания в числах
    - чтобы ожидания соответствовали реальности
    - учитывать ожидаемую полезность
- в Калибраторе можно потренироваться в интервалах

---

21. Что такое свидетельство

Конспект

- в последнем абзаце Юдковский начинает войну с религией и т.д.
- рациональное мышление → убеждения, сцепленные с реальностью - свидетельство второго уровня (для других рационалистов)
    - априорная вероятность для утверждения рационалиста заслуживает более высокий уровень
- убеждения существуют на территории
    - убеждения всех других людей - часть территории
    - мы оперируем картой, когда размышляем
    - рациональные убеждения должны быть заразительны
- введение словаря:
    - свидетельство: убеждения-о-фактах, событие, сцепленное причинно-следственной связью с тем, о чем ты хочешь узнать (с интересующей тебя сущностью)
    - сцепленность: событие проявляется по-разному для различных состояний цели
- выписать список убеждения, которые основополагающие для карьеры, общения с людьми и прогнать через сцепленность с реальностью
    - можно обнаружить висячие убеждения, которые не сцеплены с реальностью

---

22. Свидетельство: рациональное, правовое, научное

Конспект

- классификация следует практической ценности: рациональные свидетельствам имеет смысл следовать
- рациональное: убеждение, сцепленное с реальностью
- правовое (legal): преследует социальную цель: совокупность лишь конкретных типов свидетельств, следующее строгим стандартам
- научное: на его основании строим гипотезы о том, как работает мир
    - воспроизводимое знание - можно проигнорировать частное заявление и верифицировать с помощью эксперимента
    - конкретные утверждения ненаучны в отличие от обобщений
- историческое: ненаучное
- классификация не исчерпывающая
- эвристика формируется на основании свидетельств
- про подбрасывание монетки:

    - Панчон, Апофения
    - подкинутая монетка как свидетельство не сцеплена с тем, виноват ли человек для свидетельства
    - подкинуть монету для решения - способ: может быть рациональным (как способ сократить расходование ресурсов)
- применение инструментальной рациональности (выиграть в данный момент времени)

---

23. Сколько свидетельств понадобится

Конспект

- Ты пришёл к слишком сильному выводу, основываясь на недостаточном количестве свидетельств, не сумев побороть громадность пространства возможностей и априорную невероятность
- свидетельство в битах: логарифм вероятности по основанию 1/2
- для опровержения гипотезы может хватить одного свидетельства
- игра в лотерею
    - сохраняю деньги, если не играю в лотерею
    - считаем ожидаемую полезность (expected utility)
    - никакой приз не перевесит минус бесконечность потери, поэтому не следует играть
    - у денег нелинейная полезность
- как применять к прикладным задачам
    - Юдковский, Лицензия героя: оценивает вероятности по ощущениям - нужно понимать какие границы у этого метода
    - делаешь предсказания, проверяешь, строишь калибровочную структуру
    - теорема Байеса в шансовой форме
    - теорема Ферми
    - Слава: вероятностные оценки переоценены
    - Аларик: калиброваться через ставку (на что бы ты реально поставил деньги)

Менее вероятные события требует большего объема свидетельств.
Изначальная вероятность события: 1:131 115 984

Изначальное кол-во шансов	Биты информации	Кол-во шансов с битами	Вероятность правоты	Пример Юдковского
0.000000007626835184	27	1.02365649	50.58%	
0.000000007626835184	28	2.047312981	67.18%	14 ящиков

0.000000007626835184	29	4.094625961	80.37%	
0.000000007626835184	30	8.189251922	89.12%	15 ящиков
0.000000007626835184	31	16.37850384	94.25%	
0.000000007626835184	32	32.75700769	97.04%	
0.000000007626835184	33	65.51401538	98.50%	
0.000000007626835184	34	131.0280308	99.24%	17 ящиков
0.00000001	27	1.34217728	57.30%	
0.00000001	34	171.7986918	99.42%	

---

24. Самоуверенность Эйнштейна

Конспект

- В 1919 году организовали экспедицию, которая экспериментально проверяла теорию относительности Эйнштейна.
- Эйнштейн считал теорию относительности верной до подтверждающего эксперимента. Он оказался прав, но его заявление звучало дерзко.
- Традиционная Рациональность: Для того, чтобы присвоить гипотезе вероятность больше 50%, нужно 27 бит свидетельств.
- Байесовский подход: Чтобы задать гипотезу на пространстве возможных теорий, нужны свидетельства в объеме, примерно равном сложности этой гипотезы.
- Насколько это может быть правдой?
- Возможно, что у Эйнштейна уже было достаточно много (или даже больше) свидетельств истинности теории еще в процессе работы.

Вопросы

1. Пример с Эйнштейном простой и проверяемый. Однако если речь идет не о физике и не о таких глобальных теориях, как это будет работать?
2. Для особо одаренных: что такое Традиционная Рациональность? - Что входит в ТР? Например, фальсификация, необходимость прислушиваться к свидетельствам… Юдковский добавляет Бейсианство к ТР. Из-за проблемы индукции, ТР не помогает доказать теорию. С Бейсианством можно сказать: “мы уверены в гипотезе на X%” - не будет 100%, но 99.9% будет достаточно для практических задач.
3. Как можно измерить количество бит в свидетельствах?

Обсуждение

- 4 вида знания. Знаемое знание, незнаемое знание, знаемое незнание, незнаемое незнание.
- Традиционная Рациональность. https://wiki.lesswrong.com/wiki/Traditional_rationality
- Философия науки, Поппер (что делать, если в гипотезу закралась ошибка - ответа не дает) - на эту проблему дает ответ ТР
- Юдковский, “Моя дикая безбашенная юность”
- Теория относительности имеешь большой обьем - будто бы противоречит (чрезвычайные гипотезы требуют чрезвычайные доказательства) - но Эйнштейн уже обработал часть свидетельств и внедрил в текст гипотезы
- В социальных науках гипотезы постоянно меняются: например, происхождение языка https://ru.wikipedia.org/wiki/Хомский,_Ноам
- В гуманитарных науках может быть сложно поставить эксперимент из-за сложности систем - предмет для обсуждения.

О теории информации и битах (Аларик)

Теорема Байеса:

$$P(H|E) = \frac{P(H) \cdot P(E|H)}{P(E)}$$

$$P(\bar{H}|E) = \frac{P(\bar{H}) \cdot P(E|\bar{H})}{P(E)}$$

Вероятность гипотезы на вероятность обратную (априорная вероятность):

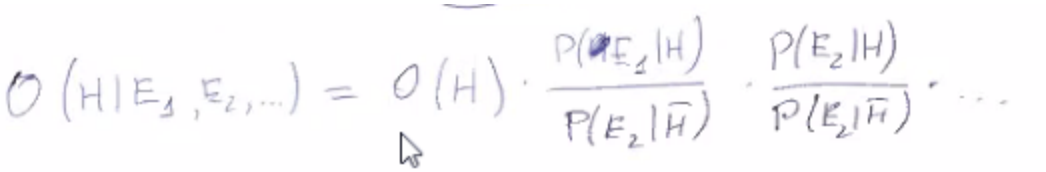
$$O(H) = \frac{P(H)}{P(\bar{H})}$$

Теорема Байеса в шансовой форме:
Например: попал ли мой друг в ДТП?
Гипотеза: он перезвонит, если попал.
P(E|H) - вероятность получить свидетельство, если гипотеза верна (друг звонит, если попал в ДТП, ожидаем приблизительно 1 от адекватного человека)
P(E|H-) - вероятность получить свидетельство, если гипотеза не верна (друг звонит, если не попал в ДТП, ожидаем приблизительно 1/10)
Например, коэфф. правдоподобия (1 делить на 1/10) = 10 меняет шансы в 10 раз.

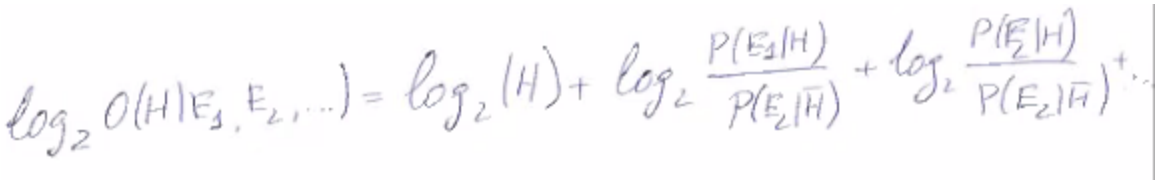
В примере из статьи о заговоре японцев:
P(E|H) < 1 - вероятность увидеть свидетельство, если заговор есть.
P(E|H-) = 1 вероятность наблюдать ничего, если заговора не существует.

Значит, наша новая вероятность будет меньше априорной.

Каждое новое свидетельство добавляет коэффициент правдоподобия:



Затем логарифмировать:



---

25. Бритва Оккама

Конспект

- Традиционная Рациональность: Чем сложнее утверждение, тем больше требуется оснований, чтобы его принять. Нужно много свидетельств.
- Чем сложнее объяснение, тем больше свидетельств необходимо, чтобы просто определить его в пространстве убеждений.
- Как можно измерить сложность объяснения?
- Как определить, сколько свидетельств потребуется?
- Бритва Оккама: Следует считать верным самое простое объяснение, не противоречащее собранным данным.
- Пример самого простого (но неверного) объяснения: “Женщина, живущая дальше по улице — ведьма, значит это сделала она” или “ЖЖНВТСО” (Хайнлайн).
- Но степень простоты не измеряется длиной предложения или количеством букв.
- Того, что теория не опровергает факты, недостаточно.
- Сложность утверждения (высказывания) - в обозначениях понятий. Пример: слово “гнев” проще (встроенное понятие), чем “дифференциальное исчисление” (не встроенное понятие). Чтобы человек понял гипотезу Тора, нужно всего лишь бросить пару фраз. Чтобы человек понял уравнения Максвелла, нужно пересказать ему несколько книг.

- Однако симулятор уравнений Максвелла написать проще, чем симулятор Тора.
- Алгоритмическая теория информации: “сложность описания” измеряется длиной кратчайшей компьютерной программы, выводящей это описание.
- Нужен оптимальный компромисс между сложностью программы и её способностью объяснять данные.
- Один бит сложности должен стоить как минимум двукратного улучшения способности объяснять.
- …
- Подвох фразы «так сделала ведьма» спрятан в слове «так». Как именно сделала ведьма?

Вопросы

1. Как на практике использовать Бритву Оккама без вычислений и сложных построений?
2. Для каких сфер деятельности полезно разбираться в индукции Соломонова и теории информации?

Обсуждение

- Упрощение гипотез в науке - в статистике предпочитаются маленькие модели, чтобы избежать оверфиттинга
- Вопрос о вероятности гипотез на примере наблюдения, что монетка падает 70:30 - есть несколько гипотез, нужно выбрать наиболее простую:
    - 1: Монетка честная
    - 2: Одна из сторон имеет большую вероятность быть сверху
    - 3: Ведьма заколдовала монетку
- Инструмент оценки простоты гипотез: минимальная длина компьютерной программы - нереалистична на практике
    - Гипотеза может описывать коротко целый класс событий
    - Мы сокращаем объем информации с помощью эмоций
    - Можно развернуть всю передаваемую информацию, как если бы мы объясняли ее с максимальной понятийной дистанцией

---

26. Сила рационалиста

Конспект

- Гость чата спросил совета по поводу недомогания и пожаловался, что врачи не оказали ему помощи.

- Автор вспомнил случаи, когда с ним происходило то же самое, и врачи были правы. Он испытывал замешательство, но проигнорировал его.
- Впихнул противоречивые данные в старую модель: “Что же, если врачи сказали “ничего страшного”, то это действительно так и есть — они бы увезли больного в госпиталь, если бы его состояние грозило бы хоть чем-нибудь серьезным”.
- Исходная история оказалась ложью.
- Мы верим случайным людям, хотя их мнение менее достоверно, чем статья в журнале (вероятно, научном?).
- Вера легче неверия; мы верим инстинктивно, но неверие требует сознательного усилия.
- Полезность модели измеряется не тем, что она может объяснить, а тем, что она объяснить не может.
- ЛИБО ТВОЯ МОДЕЛЬ НЕВЕРНА, ЛИБО ЭТА ИСТОРИЯ ЛОЖНА. - необходимо больше бит информации, чем историю гостя чата, чтобы поставить модель Юдковского под сомнение

Вопросы

1. Теоретически статья в научном журнале надежнее, чем мнение случайного человека, но на практике это часто не так. Не говоря уже о том, что у каждого журнала своя репутация.
2. Интуитивное ощущение смущения как сигнал (лакмус) тоже вызывает вопросы.
3. Проверять вообще каждое свидетельство? Это вообще возможно? И зачем? Например, огромное количество информации я получаю из интернета, например, события в других странах, наличие велосипеда в прокате, данные о материнской плате в магазине, при этом я не видела это своими глазами и опираюсь на призрачное свидетельство. Аналогично и с множеством других вещей. Как реально справиться с проблемой надежности свидетельства, опирающегося на другие свидетельства? При том, что все равно все свидетельства фильтруются через картину мира, которая нерациональна, собирает в себя не факты, а отношения, оценки и ценности?
4. В чем была ошибка Юдковского?

Обсуждение

- Модель 1: врачи скорой помощи всегда забирают для тщательного обследования - модель запрещала существование истории
- Модель 2: врачи могут отпускать людей, который, и тогда они правы
- Ошибка: Юдковский перешагнул свое замешательство и растянул модель 2 на класс историй модели 1 - нужно было остановиться на своем замешательстве и либо
- Чем больше свидетельство противоречит моей модели, тем более скептично стоит относиться к этому свидетельству

    - Пример: ЦЕРН: частицы двигались, превышая скорость света https://www.bbc.com/russian/science/2011/09/110922_collider_experiment_speedoflight - также ученым не хватило “самоуверенности” Эйнштейна
    - Пример из психологии, где ошибочная модель не была пересмотрена: установка на рост и установка на данность - эксперимент в Стенфорде должен был показать, что это странно, что установка на данность наблюдалась у измеряемых студентов
- Стоит хранить степень уверенности в модели?

---

27. Отсутствие свидетельств — свидетельство отсутствия

Конспект

- “...отсутствие (саботажа, шпионажа японцев в Америке во время Второй мировой войны) является самым зловещим... Больше чем что-либо ещё, это убеждает меня в том, что … действия Пятой Колонны будут назначены на определенное время, точно так же, как на определённое время был назначен Перл Харбор… Я считаю, что нам внушают лживое ощущение безопасности”. (Эрл Варрен)
- Теорема Байеса: Когда мы видим свидетельство, приписавшие этому свидетельству большое правдоподобие гипотезы увеличивают вероятность своей истинности за счёт гипотез, подписавших этому свидетельству меньшее правдоподобие.
    - When we see evidence, hypotheses that assigned a higher likelihood to that evidence gain probability, at the expense of hypotheses that assigned a lower likelihood to the evidence.
- Варрен: отсутствие саботажа закрепляет его убеждение о существование Пятой Колонны. Но на деле Пятой колонны вообще может не существовать. Наблюдение отсутствия саботажа увеличивает вероятность того, что Пятой Колонны не существует. (?)
- Существовавшие явления не обязательно оставляют свидетельства (например, окаменелости), но если таких свидетельств нет, то это вообще ни о чем не говорит.
- Важно быть озадаченным вымыслом больше, чем реальностью. Сила модели измеряется не тем, что она может объяснить, а тем, что она объяснить не может — только запреты могут упорядочить ожидания будущего.

Вопросы

- Различия в переводе:
    - Русский текст: “Пусть E — наблюдение отсутствия саботажа”

- Оригинал: “Let E stand for the observation of sabotage”

Обсуждение

- Гипотеза о тайном заговоре японцев на территории США -> наблюдаем отсутствие саботажа -> нельзя повышать вероятность существования организации, но нельзя и отбрасывать гипотезу

---

28. Закон сохранения ожидаемых свидетельств

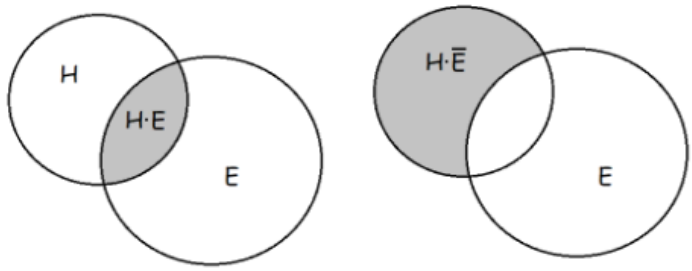
Конспект

- Духовник осужденных на смерть “ведьм” написал книгу с критикой дерева принятия решения о приговоре обвиненной в колдовстве. Например, если она грешница - то, очевидно, ведьма. Если благочестива - тоже ведьма, просто маскируется.
  Он общался со множеством таких “ведьм” и в отличие от остальных мог наблюдать все ветви этого дерева, каждая из которых только повышала убежденность в вине.
- Закон сохранения ожидаемых свидетельств: ожидаемая апостериорная вероятность с учётом будущего свидетельства должна равняться априорной вероятности.
- Для каждого свидетельства в пользу существует равное и противоположно направленное свидетельство против.
- Если имеется высокая вероятность получения слабого свидетельства в одну сторону, то она компенсируется низкой вероятностью получения сильного свидетельства в другую сторону.
- Ослабить удар опровергающего свидетельства можно лишь ослабив влияние подтверждающего свидетельства.
- Ожидание встретить свидетельство не должно само по себе сдвигать твоей убежденности. Но в зависимости от уже полученных свидетельств ты можешь сдвигаться как вверх, так и вниз по шкале уверенности. При этом нельзя ожидать, что движение будет постоянно вверх.
- Чтобы исключить возможность объяснения задним числом, важно еще до получения свидетельств формулировать что именно и куда будет тебя сдвигать
- Более подробно расписанные выкладки из эссе:

1. По определению условной вероятности: $P(H \mid E) = \frac{P(H \cdot E)}{P(E)}$

То есть:
$P(H \cdot E) = P(H \mid E) \cdot P(E)$
$P(H \cdot \overline{E}) = P(H \mid \overline{E}) \cdot P(\overline{E})$

2. $P(H) = P(H \cdot E) + P(H \cdot \overline{E})$



3. $P(E) + P(\overline{E}) = 1$

4. Далее более подробно выкладка из эссе:

$P(H) = P(H)$
$P(H) = P(H \cdot E) + P(H \cdot \overline{E})$
$P(H) = P(H \mid E) \cdot P(E) + P(H \mid \overline{E}) \cdot P(\overline{E})$

Домножаем левую часть на $P(E) + P(\overline{E})$:
$P(H) \cdot (P(E) + P(\overline{E})) = P(H \mid E) \cdot P(E) + P(H \mid \overline{E}) \cdot P(\overline{E})$
$P(H) \cdot P(E) + P(H) \cdot P(\overline{E})) = P(H \mid E) \cdot P(E) + P(H \mid \overline{E}) \cdot P(\overline{E})$

Переносим в правую часть и приводим подобные слагаемые:
$0 = (P(H \mid E) - P(H)) \cdot P(E) + (P(H \mid \overline{E}) - P(H)) \cdot P(\overline{E})$

-

Вопросы

- «Если доброта и благочестие — свидетельство того, что женщина является ведьмой, то злоба и грех должны быть свидетельством её невиновности». Есть фраза Фейнмана: «Если вы думаете, что понимаете квантовую механику, значит вы ее не понимаете». Значит ли это, что если я думаю, что не понимаю квантовую механику, то это свидетельство того, что я ее понимаю?)

Обсуждение

- Пример
  - Вероятность что есть метла и что ее нет в сумме равны 100%
  - Если у женщины есть метла - она с 20% вероятностью ведьма

- Если имеется высокая вероятность получения слабого свидетельства в одну сторону, то она компенсируется низкой вероятностью получения сильного свидетельства в другую сторону.
    - Является ли это художественным преувеличением?
    - Имеется в виду отрицание одного и того же события: высокая вероятность не встретить динозавра - низкая вероятность встретить динозавра (гипотеза, что динозавры существуют)
    - Пример, который не описывается этой формулой: найти василек алого цвета - 50 на 50 - статья про культивирование алых васильков будет сильным свидетельством в сторону “за”, статья про геном и отрицание существование алого пигмента будет сильным свидетельством в сторону “против”
        - Свидетельства не коррелируют друг с другом
    - Если есть свидетельства в пользу - должны быть и свидетельства против
    - Равное ожидание свидетельств с обеих сторон может быть, примерно одинаково свидетельства будут сдвигать нашу уверенность
    - Вероятность того, что в Мск пойдет дождь 50%. Видим тучи - сдвигает в одну сторону.
    - Случая, когда мы с высокой вероятностью может получить сильное свидетельство, не бывает
    - Свидетельства уравновешиваются, но сложным образом
- Описанные в Библии существа должны разубеждать нашу веру - звучит как-то не так?
    - Ю ссылается на прошлое эссе про японскую пятую колонну - будет странным заявить обратную вероятность: наличие саботажа должно говорить об отсутствии японской пятой колонны
    - Звучит неправильно, потому что таким образом Библия противоречит своей цели и разубеждает веру в бога
- Ожидание встретить свидетельство не должно сдвигать твои приоры - что это значит?
    - Мы находимся в равновесии до проведения эксперимента
    - После того, как он произошел, можем сдвинуть уверенность
- Вакцина - разово и навсегда меняет наше понимание мира: сложно придумать для этой цепочки.
- Индикатор - как мы будем оценивать, что мы правы:
    - Наличие свидетельств против нашей гипотезы
    - Если мы ведем себя неправильно, то приоры не меняются - мало удивлений за день происходит
    - Личная калибровка - делать предсказания в числах, в намеченный срок проверяем, сбылось или нет.
- Практика: перед тем, как наблюдать эксперимент или озвучивать свою точку зрения, спросить себя - что меня может разубедить?

---

29. Знание задним числом обесценивает науку

Конспект

- Юдковский ссылается на главу из книги Дэвида Майерса «Изучаем социальную психологию», в которой описывается эффект знания задним числом. Это тенденция преувеличивать способность предвидеть, как все произойдет, *после* того, как результат становится известным (феномен «Я знал это заранее!»)
- При этом как очевидные могут быть оценены прямо противоположные исходы, если в поле зрения попадает информация только об одном из них.
- Этот эффект не даёт заметить свидетельства, которые отличаются от того, что бы ты предсказал на самом деле.
- Знание задним числом может приводить:
    - к недооценке неожиданности научных открытий, особенно тех открытий, которые мы можем понять (например, результатов социальных экспериментов)
    - к мысли, что наука не нужна, ведь всё «и так ясно»
    - к переоценке собственных интеллектуальных способностей
    - к обвинению тех, кто принимает решения (в том числе и самого себя) за «очевидно» плохой выбор и недостаточной похвале за «очевидно» правильный выбор

Вопросы

- Знание задним числом прижилось у нас, потому что принятие любого исхода, который вписывается в здравый смысл, экономит энергию? Еще какие-то причины? - Возможно интуитивная попытка отрегулировать статус собеседника. “Ну и что, что он так сказал - я бы тоже так сказал”

Обсуждение

- Дэн Ариэли: студенты жаловались на то, что все, что он рассказывает, очевидно - взял за практику спрашивать их перед тем, как давать правильный ответ
- игра The Witness - головоломка
- Дискредитировать, обесценивать науку - как это? - Наибольшую известность имеют ученые, которые сделали контринтуитивные открытия
- Эксперимент с зефирками
- Обычно мы ожидаем некоторые исходы больше, чем другие
- Считается, что большая часть когнитивных искажений выработалась и сохранилась эволюционно, так как она была чем-то полезной - возможно, знание задним числом является интуитивной попыткой отрегулировать статус собеседника - либо для экономии ресурсов
- Практика:

    - Дешевый и прикольный способ убить скуку - предсказывать исход, затем проверять и удивляйся
    - Если вы услышали новое убеждение, перепроверь его напротив своей карты. Насколько оно соответствует твоей модели мира?
    - Напиши постик (можно личный) о том, что тебе кажется очевидным - вдруг окажется, что тебе очевидно что-то, что другим людям неочевидно - хотя, скорее всего, ты сможешь проверить свои убеждения, что будет само по себе неплохо
- Индикаторы: количество удивлений за день

О цепочке целиком

- О свидетельствах. Теория, которая описывает рациональность.
- По главам:
    - Как выглядят ожидания исходов изнутри (20)
    - Сцепленность с реальным миром (21)
    - Классификация видов свидетельств - рациональное, правовое, научное (22)
    - Решения об изменения свидетельств - экстраординарные заявления требуют экстраординарных доказатльств (23)
    - Сколько нужно для сформулирования гипотезы (24)
    - Бритва Оккама (25)
    - Способность удивляться вымыслу (26)
    - Как обновлять предсказательные силы: отсутствие свидетельств (27), закон сохранения (28)
    - Знание задним числом (29)
- Вообще:
    - Числовые законы, зачем нужна математика
    - Рациональность
    - Маленькие ощущения изнутри: что именно чувствую, когда ожидаю исхода? Как выглядит уверенность?
- Оффтоп - как планировать исход с будущем:
    - Думать от будущего состояния: представить идеальный вариант и расписать идеальные условия для него
    - Думать от status quo: дерево принятия решения: связный граф - для грубой оценки вероятностей
        - Интуитивные оценки разных параметров: время, деньги, позитивный реинфорсинг
        - Двухуровневые или трехуровневые деревья - это ок
        - Оценить с помощью внкутренней функции полезности
        - Измерить дистанцию между листочками и идеальным исходом
        - Если найден листочек, который ведет ко всем идеальным исходам - это доминантная стратегия

        - Правильно думать: что нужно сделать, чтобы все закончилось так, как я хочу?
    - Идея Юдковского: прилагай экстраординарные усилия

Загадочные ответы

30. Лжеобъяснения

Конспект

- Учительница нагрела пластину перед обогревателем, а потом развернула её. Сторона у обогревателя стала холоднее, чем обратная сторона.
- На вопрос “почему так?” студенты, вместо того чтобы испытать замешательство, пытались придумать объяснение, которое впишет неожиданное событие в привычную теорию.
- В таком случае использование научных терминов для объяснения ничем не лучше объяснения “это магия!” - оно также никак не ограничивает пространство возможных событий. Однако наличие “научного” объяснения создает иллюзию, что ответ на вопрос действительно найден.

Вопросы

Обсуждение

- Мы продолжаем пытаться подбирать пароли, будучи взрослыми
- Задачка про трех черепашек - иллюзия прозрачности
- Когда перед тобой ставят задачу, нужно попробовать уйти в мету - зачем она ставит такой вопрос?
- Практика: проверяй, сможет ли твоя гипотеза объяснить противоположный результат. Если да, вероятно что-то не так с гипотезой.
- Научно-популярная литература передает знания, художественная - опыт

---

31. Угадай слово, загаданное учителем

Конспект

- Юный Юдковский долго полагал, что знает смысл слов: свет — это волны, звук — это волны, материя — это волны. После знакомства с фейнмановскими лекциями по физике он понял, что на самом деле не знал.
- Казалось бы если фраза «Свет — это волны» истинна в устах физика, то и в устах любого другого она тоже истинна.
- Но у слов нет встроенных значений. Когда я слышу слоги «бо-бёр», в моём мозгу возникает образ большого грызуна; но это факт о состоянии моего разума, а не о слогах «бо-бёр»
- Внутри разума может быть связь между фразами «Свет — это волны» или «из-за теплопроводности» и какими-нибудь гипотезами, но эти фразы, сама по себе, не являются ни истинными, ни ложными.
- Но в школе за правильными паролями следует положительное подкрепление в виде оценок, и это закрепляет иллюзию, что у фразы есть истинное значение.
- Фраза “из-за теплопроводности” обретает смысл если:
    - она ограничивает ожидание будущих переживаний
    - ты предпримешь попытки проверить её, а не ограничишься лишь попыткой угадать пароль.
- Слова не считаются, а имеют значение лишь контроллеры ожиданий

Вопросы

- Представьте, что вас перенесли в древнюю грецию. Какими знаниями вы сможете поделиться с греками (предположим, языковой барьер вы преодолели)?

Обсуждение

- Критика системы образования
- Медитация, Випасана: истинное знание заключается в том, чтобы пропустить его через собственный опыт
- Как определить, если у тебя есть знания или если их нет
    - Например, можно объяснить незнающему человеку свои знания - сразу будут заметны свои пробелы
    - Сложно определить, насколько ты компетентен
    - Если есть модель, которая отражает реальность
    - Лучше знать, чтобы приспособиться к изменяющейся реальности
    - Успешность в карьере не является критерием обладания знаний - ученые тоже допускают глупые ошибки и могут иметь когнитивные искажения
- Про контроллеры ожиданий: я не знаю, как работает теплопроводность, но я ожидаю, что она работает именно так
- Насим Талеб, заблуждение про “зеленый лес”: когда детали заслоняют глобальную картину - часто не нужно понимать, как что-то работает, чтобы разбираться в вопросе

    - Проблема в том, что некоторые знания ненужные для требуемой цели: т.е. характеристика леса для продажи
    - Трейдер работал, угадывая ответ - не понимал, почему так происходит
- В рац-сообществах по умолчанию ожидается желание лучше понимать мир. Не знать, как работать тот или иной кусок реальности - это ок.
    - Научная картина мысли дает более гибкую и надежную основу для нестандартных ситуаций
    - Нужно понимать ограничения своих возможностей - полную модель реальности построить нельзя
- Канеман: в каких областях работает интуиция - важно получение обратной связи
- Есть “знания” (пароли, заученные в школе и т. д.), на основаниях которых как бы можно предсказывать события, но результат будет угадайкой

---

32. Наука как одеяние

Конспект

- В Людях Х говорится: «В каждом человеке… есть генетический код… вызывающий мутации». Благодаря этому можно приобрести полезные способности, например,метать молнии.
- Но факт получения сложных биологических приспособлений, отвечающих за эту способность, за одно поколение, в результате мутации, одним махом вдребезги бы опровергло неодарвинистскую модель естественного отбора.
- Научные термины используются в подобных случаях лишь для создания атмосферы наукообразия. Это никак не связано с ограничением ожиданий
- Термины растягиваются до объяснения чего угодно, вплоть до явлений из сфер, которые к этим терминам не относятся (например, эволюционную биологию приплетают к развитию информационных технологий)
- Использование научных терминов - способ выглядеть причастным к научной тусовке. Т.е. частный случай убеждений-одеяний
- Ну и еще Юдковский бомбит о том, что такое представление науки в массовой культуре:
    - превращает науку в литературный жанр
    - приводит к тому, что к серьезным вопросам относятся несерьезно (например, ассоциируют сверхчеловеческий ИИ с литературным жанром беллетристики о конце света)

Вопросы

- Вопросы к себе:
    - Есть ли что-то такое в науке, верой во что вы гордитесь, но до сих пор не применяете вашу веру на практике?

    - Какие возможные варианты будущего ваша вера запрещает?

Обсуждение

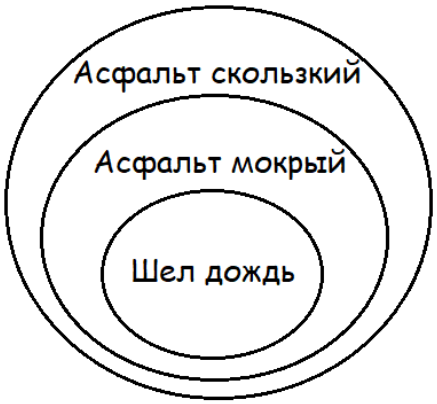
- Примеры:
    - Люди Икс - мифология науки:
        - Хорошо: стимулирует детей становиться учеными
        - Плохо: наука используется для описания магии - ей могут перестать верить в результате
        - Проблема с паттерном науки в общественном сознании
        - По контексту должно быть понятно, что Люди Икс - вымысел
    - Паблик ВК “Разумный замысел” - хотелось бы верить без предпроверки
    - Сериал “Космос”
- Ненаучное стало немейнстримом - его стали упаковывать в научную терминологию
- Хуже, когда фальшивку выдают за знания, упаковывая в научную лексику, чем ненаучное априори
- Когда что-то происходит “магическим образом”
- Цепочки:
    - Внутренняя направленность на человека, который их читает - первый враг это мы сами
    - Направленность на общество
- Наука как одеяние: объяснение закономерностей научным языком

---

33. Лжепричинность

Конспект

- Алхимики объясняли огнь при помощи термина «флогистон». Флогистон содержался в веществах и покидал их в процессе горения. Пепел, оставшийся после сгорания, считался истинным материалом веществ. Огонь в закрытом сосуде быстро потухал, потому что воздух насыщался флогистоном и больше не мог его вместить. Все явления, с горением, сначала наблюдались, а потом вписывались в теорию.
- Специалисты предполагают, что люди думают о причинно-следственных связях, используя нечто вроде направленных ациклических графов или байесовских сетей. Поскольку шел дождь, тротуар мокрый; поскольку тротуар мокрый, он скользкий



- Если тротуар скользкий, то это свидетельство в пользу дождя (и мы можем посчитать насколько сильное). Однако если мы уже знаем о том, что тротуар мокрый, то скользкость уже не даст новой информации о дожде.
- Нельзя возвращать информацию в узел, в котором она возникла. Если мы узнали что-то о том, что тротуар мокрый, мы можем предполагать, что идёт дождь. Но это предположение о дожде не может увеличить убежденность в том, что тротуар мокрый.
- А с флогистоном происходило именно так. Алхимики получили информацию о том, что огонь горячий, обновляли знание о флогистоне, а после этого утверждали, что огонь горячий из-за флогистона

Вопросы

Обсуждение

- Флогистон не контролирует ожидания
- Флогистон может кое-что предсказывать - что зола не будет гореть - Юдковский не совсем прав, что этот пример подходит
    - Основа не горит, флогистон горит
- Теплород и флогистон: теплород был ближе к правде, флогистон предсказал отрицательную массу, потом оказалось, что это было заблуждением
- Двойной учет свидетельств - пример из Терапии настроении Бернса

---

34. Семантические стоп-сигналы

Конспект

- Затычка вроде «Это Бог!» - это не столько сознательное утверждение, сколько дорожный знак на трассе для мыслей, говорящий «дальше не думай, проезд закрыт». Такая затычка не разрешает вопрос, а останавливает цепочку естественных вопросов и ответов.
- Такими затычками могут быть «демократия!», «корпорации!» и много чего ещё
- Понятие «семантический стоп-сигнал» нельзя превращать в обобщённый контраргумент против вещей, которые тебе не по душе («Да ну, это просто бессмыслица, приправленная семантическими стоп-сигналами!»). Слово не может быть стоп-сигналом само по себе; вопрос заключается в том, производит ли оно этот эффект на конкретного человека.
- Главная черта семантического стоп-сигнала — нежелание рассмотреть следующий очевидный вопрос.

Вопросы

- Почему парадокс первопричины кажущийся?

Обсуждение

- Thought-terminating cliches: по смыслу похоже на стоп-сигналы, но также несут информацию (что более подробную информацию получить невозможно) - стоп-сигнал не предполагает конец диалога
    - “Пути Господни неисповедимы”
- Стоп-сигнал может указать на разницу в понимании мира - например, на понятийное расстояние
    - “Потому что так надо”
- Из контекста эпистемологии: стоп-сигнал может упираться в терминальную ценность собеседника
- Про феминизм и феминитивы: как только начнешь в диалоге с феминистом критиковать феминитивы, собеседник триггерится и обрушивает на тебя “праведный гнев”
    - Связано с запретными темами
    - Похоже на кейс “соломенного чучела” со стороны собеседника, который цепляется за часть твоего аргумента, чтобы полностью аргумент разрушить
    - Реакция на феминитивы больше к статье "убеждение как одеяние", про нездоровое отношение к критике.

---

35. Таинственные ответы на таинственные вопросы

Конспект

- Витализм: загадочные различия между живой и неживой материей могут быть объяснены посредством «жизненной силы»
- Теория «жизненной силы» также как и флогистон обогащалась только за счёт уже наблюдаемых явлений и не ограничивает ожидания. «Жизненная сила» также играет роль семантического стоп-сигнала.
- Но еще одна черта подобного рода загадочных ответов — преклонение перед таинственностью. Когда вместо того, чтобы пытаться объяснить сложное явление, ему приписывается принципиальная необъяснимость и непохожесть на всё, что было ранее.
- Однако непонимание, замешательство и невежество — это то, что рисуется на карте, а не то, что можно обнаружить, гуляя по местности. Не существует явлений таинственных самих по себе Ошибка здесь в допущении того, что ответ может быть таинственным. Ответ должен хотя бы немного снижать уровень таинственности, иначе это не ответ, а затычка для любопытства.
- Признаки таинственного ответа:
    - не контролирует ожидания
    - модель представляет из себя сплошную субстанцию, а не механизм, состоящий из частей
    - противопоставление науке; создание уникального класса для этого явления
    - после ответа уровень таинственности не снизился

Вопросы

- В чем отличие от флогистона, эмерджентности, сложности? Важно ли это различие?

Обсуждение

- Лорд Кельвин крепко держался за витализм, когда всю его сознательную жизнь уже исследовательски это было опровергнуто

---

36. Тщетность эмерджентности

Конспект

- Эмерджентность — это то, как сложные системы и паттерны появляются из множества относительно простых взаимодействий

- Те же претензии, что к флогистону и витализму. Выступает как стоп-сигнал, не добавляет новой информации о явлении, не позволяет сделать предсказание. Не является механизмом, а используется как объяснение само по себе. Содержит священную тайну для поклонения.
- Те, кто предлагают гипотезу «эмерджентности», признаются в своем незнании внутреннего устройства и гордятся этим; они противопоставляют «эмерджентные» науки и «обычные»
- Совет: попробовать заменить эмерджентность (или другое понятие, которое одним махом объясняет всё) на магию:
    - жизнь — эмерджентный феномен -> жизнь — магический феномен.

  Каждое утверждение дает одинаковый объем информации о поведении феномена. Каждая гипотеза подходит под одинаковый набор результатов.

Вопросы

- В чем отличие от флогистона, сложности, витализма? Важно ли это различие?

Обсуждение

- Эмерджентность по Юдковскому - синоним определенного вида магии - в которой целое больше суммы отдельных частей
- Например, навешивать на компьютер новые свойства, чтобы объяснить ИИ - не можем объяснить, почему навешивание доп свойств приведет к тому, что у компьютера появится самосознание
- Хайлайн, Луна - суровая хозяйка: идея эмерджентности
- Определение из теории систем: часы - пример системы, которая обладает эмерджентностью

---

37. Скажи нет «сложности»

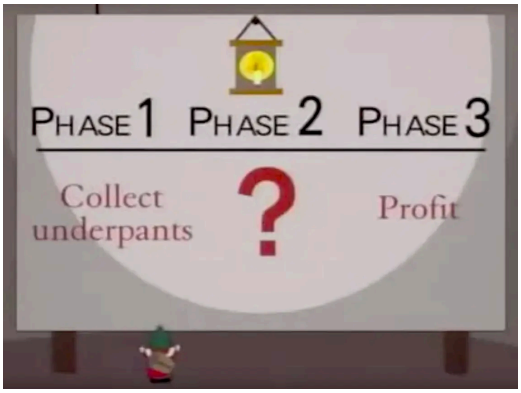
Конспект

- Сложность не должна быть самоцелью. Продумывая дополнительный элемент теории, ты увеличишь сложность, но делаешь это не ради добавления сложности.
- Фраза «X должно быть достаточно сложным, чтобы получить Y» не делает Y достижимым, не говорит о том, что именно нам нужно сделать с X.
- При этом, термин «сложность» не бесполезен. Он нужен, чтобы описать уже продуманную модель или понять насколько сложна гипотеза и как она соотносится с имеющимися свидетельствами. Но сам по себе он не объясняет как устроен какой-то кусок территории. "Сложность" не должна становиться затычкой мест, где ты не разобрался.

- Понимание, что ты ещё не закончил важно. По-умолчанию мозг стремится избавиться от неопределенности и незавершенности.
- Помимо этого, говорить "тут я ещё не разобрался" может не позволить статус.
- Ещё можно не решаться исследовать неизведанное, потому что страшно потерять время впустую.
- Полезно называть непонятные моменты «магией», а не каким-то умным словом. Чтобы впоследствии не замять этот не понятый кусок.

Вопросы

- Контрпример из South Park’a:



- В чем отличие от флогистона, витализма, эмерджентности? Важно ли это различие?

Обсуждение

35 – 37 вместе

- Флогистон, сложность, эмерджентность: явления одного класса - несостоятельные объяснения

---

38. Подтверждающее искажение: взгляд во тьму

Конспект

- Испытуемые, которым предлагалась задача «2-4-6», обычно пытались придумать положительные примеры, а не негативные — они применяли гипотетическое правило для создания триплета, а потом смотрели, будет ли он отмечен как «да»
- Испытуемые не пытались придумать контрпримеры к своей гипотезе

- Про эффект подтверждающего искажения можно знать, но не замечать когда он срабатывает. Нужно еще уметь еще до вербального формулирования думать о негативном примере, а не о положительном
- Связь с предыдущими эссе — нужно найти исход, который невозможен при истинности гипотезы, и проверить действительно ли он не возможен

Вопросы

- Ценность блога Юдковского заключается в том, какую часть мира он не способен объяснить?

Обсуждение

- Практика:
    - Английская википедия, раздел Критика
    - Steelmanning
    - Уличная эпистемология
    - Диалог: представить себе человека, к которому ты хотелась бы подкопаться
    - Пример с Малфоем, который представлял себя доктором, который представляет себя доктором, который хочет запретить публикацию статьи Поттер про чистоту крови
    - Помогает делать предсказания
    - Помогает фиксировать, записывать мысли
- Проблема: прививка от аргумента - упорно не принимает точку зрения

---

39. Закономерная неопределённость

Конспект

- Эксперимент Тверского: ошибочное интуитивное поведение участников эксперимента, которые выбирали смешанную стратегию ставок (правы в 58% случаев) вместо того, чтобы постоянно ставить на синий цвет (правы в 70% случаев)
- Испытуемые продолжали выбирать красный, как будто они полагали, что способны предугадать случайную последовательность
- Юдковский предполагает, что стратегия «всегда-на-синее» просто не приходила испытуемым в голову.
- Возможно им казалось, что территория содержит случайность, а значит и стратегия должна содержать случайность. Однако случайное поведение не решает проблему.

- Оптимальная стратегия поведения в неопределенной ситуации: нельзя выбрасывать рациональность при встрече с нерациональностью
- “Ты не гасишь огонь огнем, ты гасишь огонь водой. Но эта мысль подразумевает лишний шаг, новый концепт, не активируемый напрямую формулировкой задачи.”

Вопросы

Обсуждение

- Что если попробовать провести эксперимент с другом?
- CFAR - центр рациональности - именно этот пример в контексте человеческих отношений
    - Собеседник действует в добрых или злых намерениях
    - В среднем чаще в добрых
    - Оптимально будет все время вести себя все время, предполагая что он добрый всегда
- Пример в камень ножницы бумага - если вести себя случайно, то будешь побеждать
    - Есть просчитанные ходы, которые чаще люди используют
    - Пока не определил стратегию собеседника, лучше играть случайно
- Пример игры
    - Два игрока
    - Один пишет на двух карточках по разному целому числу
    - Другой берет карту и должен угадать, если это число большее или меньшее
    - Угадывает - получает рубль
    - Бесконечное количество раундов
    - Оптимальная стратегия - отгадывающему нужен генератор случайных чисел - вероятность победы чуть больше ½
- Какую стратегию выбирать: та что с большей вероятностью приводит к успеху
- Пример: ходить в ресторан - в тот же самый или в новый?
- Выбор геймдизайнера на работу, неограниченное время
- Логично != интуитивно

---

40. Моя дикая и безбашенная юность

Конспект

- Традиционная Рациональность и Путь Байеса
- ТР: ретроспективные предсказания не требуются, нужно сделать фальсифицируемое предсказание, отбросить гипотезу в случае фальсификации

- ТР позволяет делать гипотезы, не фокусируя уверенность на каком-то из исходов. В ТР тебе позволено гадать и тестировать свои догадки. Так можно потратить много времени, перебирая кучу фальсифицируемых, но бесполезных гипотез.
- Путь Байеса: основан на математике и фундаментальных правилах поведения человека
- Молодой Элиезер следовал ТР и верил в Таинственный Ответ
- Поздний Элиезер:
    - Почему рационалисты не правят миром?
    - Фальсификация по ТР занимает кучу времени и ведет лишь к новым интересным ошибкам, ПБ же позволяет получать правильные ответы
- Совет начинающим рационалистам: «Не пытайтесь строить сложных цепочек рассуждений и планов». Единственно верная сложность - та, где каждый шаг описан с безжалостной точностью

Вопросы

Обсуждение

- Стоит выдвигать гипотезы после сфокусирования ожиданий
- Нет ретроспективных свидетельств -> вероятность встретить динозавра 50:50
- Выбор гипотезы: пример поиска убийцы - предположим, что это Джон Джонс - можно фальсифицировать - хотя априорная вероятность не высока (одинаковый уровень вероятности, что каждый из жителей может быть убийцей)
- ТР: https://wiki.lesswrong.com/wiki/Traditional_rationality
    - Фейнман
    - Кун
    - Поппер
- Каким способом я фальсифицирую? Каким способом я быстрее всего сдвину свою вероятность? ТР подвержен предвзятости убеждения. Нужно понять, какие свидетельства повлияют.
- Имеет смысл искать, почему ты веришь в что-то
- Интуиция основана на опыте
- Лучше доверять интуиции эксперта

---

41. Неспособность учиться у истории

Конспект

- Получение ответа должно делать вопрос менее запутанным.

- Мы не обращаем внимание на то, что вещи, обыденные сейчас, были Таинственными Вопросами ранее - мы их решили без Таинственных Ответов
- “Мы изучаем историю, но мы не живем ей, не переживаем этот опыт.”

Вопросы

- Отсылка к ошибке хайндсайта (суждение задним числом)?

Обсуждение

---

42. Делая историю доступной

Конспект

- Обобщение на основе вымышленной истории - когда в обсуждении ИИ делают отсылки к “Терминатору”. А в общем виде - это использование вымышленной истории как основу для формулирования и подтверждения гипотезы.
- Обратная ошибка: игнорирование исторических свидетельств, потому что мы лично не застали время, когда они происходили
- Ошибка наших предшественников: таинственные вопросы (почему камень неживой) “решались” таинственным ответом (витализм)
- Чтобы действительно ощутить силу истории, ты должен думать о событиях так, как если бы они случились с тобой, пересилить ложную амнезию, вызванную рождением в конкретную эпоху.
- Элиезер мысленно переживает ход истории: “Вспомни, как ты был рожден в племени охотников-собирателей тысячи лет назад, когда еще никто не знал о Науке. Вспомни, как ты был шокирован до глубины души, когда Наука объяснила великие и ужасные священные тайны, которые ты так восхвалял. Вспомни, как ты думал, что сможешь летать, если съешь нужный гриб, как ты разочарованно усвоил, что никогда не сможешь полететь, а потом полетел. Вспомни, как ты ты всегда думал, что рабство - это правильно и хорошо, а потом передумал. Не надо воображать, как бы ты мог предсказать перемены - ведь это проявление амнезии. Вспомни, что на самом деле ты не угадал. Вспомни, как век за веком мир менялся так, как ты и представить не мог.”

Вопросы

- Элиезер описывает Путь Байеса?

Обсуждение

---

43. Объяснить, поклониться, пренебречь

Конспект

- Трилемма: объяснить, поклониться, пренебречь
- Рационалисты считают своим долгом постоянно нажимать «объяснить», но это выглядит, как дорога без конца. Каждое такое нажатие вызывает новый вопрос и новую возможность выбрать: объяснить, поклониться или пренебречь.
- Поклониться - это придумать таинственный ответ на таинственный вопрос
- Юдковский выбирает поклониться Объяснению, Не Требующему Объяснения, но также может ее продолжить вопросом «Хм, это какой-то парадокс. Интересно, как он разрешается?»
- Если весь вопрос кажется тебе неважным, или неуместным, или тебе кажется, что лучше подумать о нём потом — значит, ты нажал «пренебречь».

Вопросы

Обсуждение

- Стоит делать выбор эксплицитно
- Всегда что-то будет необъяснимо - что мы не хотим обсуждатть
- Кнопка “поклониться” - превознести факт отсутствия ответа (Необъяснимо!)
    - Может быть личная выгода в выборе “поклониться”
- Пример: сингулярность - черная дыра - нельзя объяснить нынешними законами физики
    - Новые открытия в физике могут объяснить
    - В конце будет лишь одно истинное объяснение - физика это хороший претендент на подобное
- Некоторые ответы не важны или невозможны (наш бюджет на поиск ответов ограничен)
    - “You know, modern computer science gives us lots of examples of questions that we can't ever know the answer to even though they have mundane answers. These could require halting oracles to answer, but could also simply need physically unrealizable computing power due to their complexity class. Maybe science ends when the next step in the causal chain is simply probably not answerable with realistic resources”
- Набор кнопок не исчерпывающий

- Пример: эволюционизм

---

44. «Наука» — затычка для любопытства

Конспект

- Окружающий мир наполнен вещами, на которые другие люди (ученые, биологи, учителя и т.д.) подготовили для нас объяснения
- Любопытство гаснет как только становится известно, что тайна, на которую вы смотрите, уже кем-то раскрыта
- Объяснение словом “наука” не оставляет последующих вопросов - собеседник готов поверить в этот ответ (в отличие от объяснения вроде “магия”)

Вопросы

- Как направлять свое любопытство? Есть важные темы, в которых любопытство помогает побеждать (упорно искать ответ своими силами, игнорируя тот, что нам предоставили другие люди). Есть малозначимые темы (я смотрю на экран компьютера, текст появляется при нажатии клавиш - есть простое объяснение - это техника - но можно включить любопытство и разобраться). Время ограничено.

Обсуждение

- Любопытство - когда что-то удивляет
- Кто-то другой знает ответ
- Зеленый слоник: https://godescalc.wordpress.com/2012/06/24/overlooked-elephant/
- Хватит жаловаться на то, что наука якобы уже все открыла
- Полезно ли интересоваться вещами, ответы на которые уже известны?
- Любопытство притягивают вещи, которые мистическим образом “необъяснимы” - например, живой огонь в Иерусалиме
- Совет: не принимать ответ собеседника автоматически
    - Например, не принимать сходу мнение о том, что коронавирус вообще не опасен - было бы полезно на момент начала пандемии
- https://lesswrong.ru/w/Выученное_непонимание

---

45. Поистине часть тебя

Конспект

- Убеждение настолько твоё, насколько оно соединено с другими знаниями, чувственным опытом, ожиданиями, действиями, которые ты осуществляешь в связи с этим убеждением.
- Если ты видел бобра, грызущего дерево, ты скорее всего сможешь опознать его в будущем, независимо от того, что это существо называется «бобер». Но если ты только слышал факт о том, что бобры грызут деревья, то возможно не опознаешь бобра при встрече.
- Способ проверить можно ли доверять имеющемуся знанию - задаться вопросом «смогу ли я восстановить это знание, если оно почему-то исчезнет из моей головы?»
- Если ты не можешь что-то создать сам, ты не сможешь это и восстановить, если ситуация этого потребует
- Витгенштейн: «Колесо, которое можно свободно вращать, не задевая других частей, не является частью механизма».
- Если мысль первоначально пришла извне, убедись, что она также исходит изнутри
- Полезные вопросы себе:
    - Как бы я смог восстановить это знание, если оно будет удалено из моего разума?
    - Как много своих знаний я смогу восстановить самостоятельно после удаления?
    - Какова допустимая глубина удаления?
- Пост Искусственное суммирование: представь мир, в котором процесс сложения симулируется конечным набором процессов вроде “Сложить(Семь, Шесть) = Тринадцать”

Вопросы

- Почему знание, которое ИИ не сможет восстановить сам — чудовищная опасность?

Комментарии

- Проблема ИИ - создать интеллект с нуля, который сможет выполнять задачи, которые может выполнять человек, или даже больше. Создание начинается с нуля - необходимо представить, что изначально машина не обладает никаким знанием. Сложность в том, чтобы деконструировать большие процессы мышления человеком (думать, понимать, чувствовать) до мелких деталей, которые их составляют. Без заложенных внутрь процессов слово “понимать” не обладает смыслом, это лишь удобное обобщение.

Обсуждение

- Можно ли восстановить знание?
- На что похож пример в статье: направленный граф, семантическая сеть
- Как расширить термин счастья: для компьютера нет разницы, есть название “Happiness” или “G1073”

- Доклад с международной встречи LW: предсказание = сжатие: https://www.lesswrong.com/posts/hAvGi9YAPZAnnjZNY/prediction-compression-transcript-1
- Сюжет в художественной литературе: возврат в прошлое и попытка воссоздать знание, известное в настоящем
- Эрик Флинт, научная фантастика, “1632”
- Нельзя воссоздать всю математику - нужны аксиомы
    - Однако Евклид не строил геометрию абстрактно - он искал решение практическим задачам
    - Лобачевский поменял аксиомы для конкретных нужд
    - Можно представить возможность воссоздать математику для практических задач

McDermott 1976, “Artificial Intelligence Meets Natural Stupidity” (контекст из Поистине часть тебя)

- Статья: https://dl.acm.org/doi/10.1145/1045339.1045340
- Лисп: язык программирования - создан для работ по искусственному интеллекту и до сих пор остаётся одним из основных инструментальных средств в данной области
    - Выражения представляются списками в префиксной записи: первый элемент должен быть формой, то есть функцией, оператором, макросом или специальным оператором; прочие элементы — аргументы этой формы, передаваемые форме для обработки. Арифметические операторы записываются по тому же принципу, например (+ 4 (* 2 3)) выдаёт 10
- Автор: область ИИ критикует - она на грани между наукой и шарлатанством/чудачеством
- Автор критикует многообещающие мнемоники, используемые в языке и дающие преувеличенное представление о процессах
    - Концепты, которые берут из человеческого языка, и нагруженные смыслом
    - Пример: программист называет главную петлю в своей программе UNDERSTAND - это абстрактный термин, который не описывает, что именно происходит
    - Также термин GOAL
    - Другой пример: GPS: Generalized Problem Solver - универсальный решатель задач: компьютерная программа, предназначенная для работы в качестве универсальной машины для решения задач, но способная решать лишь простые задачи
- Ветвь IS-A безобидна на первый взгляд
    - IS-A соединяет два узла в “семантическую сеть” (неправильное название)
    - Объяснение примера с IS-A
- Недостаточно просто переписать связь в человеческом языке теми же словами и назвать это программированием

О цепочке целиком

- Про какие несостоятельные ответы мы поговорили:
    - Попытка подобрать пароль или магическое слово
    - Наука как одеяние
    - Семантический стоп-сигнал
    - Таинственный ответ
    - Эмерджентность
    - Сложность
    - Флогистон
- Обобщает случаи того, когда мы отказываемся искать ответ
- Рассмотрены некоторые варианты загадочных ответов:
    - Попытка угадать пароль или магическое слово (потому что поощряют в школе и не только)
    - Использование науки как одеяния (потому что наука это показатель качества)
    - Флогистон, как пример нарушения причинно-следственных связей
    - Семантические стоп-сигналы (ответ, который останавливает дальнейшие размышления)
    - Витализм как пример таинственного ответа на таинственный вопрос (преклонение перед тайной)
    - Эмерджентность - пример современного таинственного ответа
- Все эти ответы объединяет одно - отсутствуют контроллеры ожиданий. А следовательно невозможность узнавать действительно ли теория хорошо описывает реальность
- Далее описаны факторы, которые способствуют появлению загадочных ответов и их закреплению на карте:
    - Подтверждающее искажение и склонность тестировать положительные примеры
    - Попытки справиться с хаосом при помощи хаоса и неспособность видеть простые закономерности
    - Использование лишь TP, но не байесианства. Как следствие трата времени на проверяемые, но несостоятельные гипотезы.
    - Использование вымышленных свидетельств и наоборот неспособность учитывать исторические свидетельства. Отсутствие опыта преодоления таинственности.
    - Желание всегда объяснять сопряжено с неопределенностью (приходится учиться её терпеть)
    - Снижение любопытства к объясненным обыденным вещам

- Убеждения, плохо встроенные в систему убеждений. Убеждения со слабыми связями.

---

46. Простая истина

Конспект

- Истина не обязана быть сложной - если очевидный ответ, который первым приходит в голову, остается очевидным после раздумий, следовать ему будет разумно
- Литания Тарского
- История:
  - Контекст: мир без математики
  - Автор: пастух, который изобрел способ считать овец, бросая камушки в ведро
  - Отри: подмастерье, которого пастух научил способу, но который сам не понимает его принцип работы
  - Марк: сенатор, спорит с Отри, пытаясь выяснить, каким магическим способом работает ведро с камушками
  - Инспектор Дарвин: разрешает спор Отри и Марка экспериментом, в котором каждый может сохранить свою истину - позволяет Марку проверить его истину применительно к реальности
- Темы обсуждения:
  - Марк: истина субъективна, не существует реальности, есть лишь убеждения, которые ее описывают
  - Отри: субъективная вера не может изменить реальность
  - Пастух: метод ведра с камушками работает независимо от того, верим мы в него или нет
- “Я не могу создать собственную реальность в лаборатории, поэтому понять, что это такое, пока нельзя. Но время от времени я сильно верю, что что-то случится, а вместо этого затем происходит что-то другое. Мне нужно как-то называть это «что бы то ни было», которое определяет мои опытные результаты, поэтому я называю это «реальностью». Эта «реальность» как-то совсем не совпадает иногда даже с моими лучшими гипотезами. В некоторых случаях, когда у меня есть простая гипотеза, которая прекрасно согласуется со всеми известными мне данными, даже и тогда случаются сюрпризы. Так что мне нужно по-разному называть те штуки, которые определяют предсказанные мной результаты, и ту штуку, которая определяет опытный результат. Первое я называю «верой», а второе — «реальностью».”

Комментарии

- Карта будет отражать территорию, если есть процесс, контролируемый реальностью, который помогает рисовать карту
    - Примеры такого процесса: механизм бросания камушков в ведро, человеческий глаз, смотрящий на шнурки
    - Антипримеры: эссенциализм, магические палочки
- Откуда мы знаем, что кроме карты (убеждений) есть территория (реальность)? - Когда мы делаем предсказание с уверенностью, а оно не оправдывается
- Истина: совпадение между верой и реальностью
- Прыжок с обрыва показывает разницу между истинными и сильными убеждениями
- Есть ли способ сравнить убеждения из своей головы с реальностью или мы лишь умеем сравнивать убеждения с убеждениями?

Вопросы

- Насколько полезно понимать траекторию спора Отри и Марка? Что насчет размышлений на мета-уровне?
- Можно ли идентифицировать темы тома “Карта и территория” в отдельных моментах спора Отри и Марка?

Прочие ссылки

- Тред с обсуждением поста: https://news.ycombinator.com/item?id=4960992
- Обновленная идея Юдковского про истину: пост Полезная идея истины

Обсуждение

- Комментарии:
    - Убеждение как одеяние
    - Отсылки к разным способам аргументации
    - Разумные вопросы Марка
    - Пастух придерживается ТР
    - Как можно согласиться и не соглашаться
    - Истина есть
    - Ведро с камушками - метафора арифметики
    - Цикл в программировании
- Что вытащить для практического применения?
    - Белый шум? - Формат неподходящий
    - Когда пытаешься соотнести с реальной жизнью - голова закипает
    - Мало инструментальной рациональности в этом томе
    - ПЯ некорректрна для практ. применения
- Итог: Воспринимай ведро с камушками как ведро с камушками

О первом томе целиком

- Словарь: что такое рациональность и зачем, концепция карты и территории, краткий обзор когнитивных искажений.
- Надежные убеждения ограничивают ожидания. Полезно задаваться вопросом: какой опыт предсказывается моим убеждением, а какой запрещается?
- Словарь: что такое свидетельство, сцепленность, сила свидетельства, сложность объяснений, бритва Оккама.
- Сила рационалиста - способность быть озадаченным вымыслом больше, чем реальностью.
- Недостатки традиционной рациональности. Преимущества байесианства.
- Случаи, в которых мы отказываемся искать ответ. Как вышло, что мы удовлетворяемся ответами, которые не делают мир понятнее.

Обсуждение

- Убеждения должны ограничивать ожидания
- Делать предсказания в реальности
    - Калиброваться
    - Спорить на деньги
    - Писать публично о предсказании
    - Обязательно записывать
    - https://www.lesswrong.com/posts/AM5JiWfmbAytmBq82/betting-with-mandatory-post-mortem
- Прогнозом не всегда можно объяснить реальность
- Как перенести в бытовую сферу
- Часто на интуиции
- Дракон на балконе
- Интуиция тоже работает (https://lesswrong.ru/w/16_видов_полезных_предсказаний), можно ее тренировать
- Рациональность не про точность - можно обойтись системой 1. Система 2 - для точности.
