

Эссе с LessWrong

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 первого тома “Рациональности: от ИИ до зомби”:

<https://docs.google.com/document/d/1Yil17JLO5jWvAj2E7W7BKA6TCQzIQZnOWdl9zyb7bv4/edit?usp=sharing>

Ссылка на конспект второго тома “Рациональности: от ИИ до зомби”:

<https://docs.google.com/document/d/1eH5THY75tvMwQ7vBrL5Dx53Of2whPRoLSqKRP7Gq2io/edit?usp=sharing>

Скотт Александер, О пользе любезности, общин и цивилизации (17.08.2020)

О чем это эссе: ответ Скотта Александера на критику от Артура Чу, победителя телевикторины Jeopardy!

Чу:

- LessWrong восхваляет принцип “рациональность побеждает”, проигрывая на самом деле
- Для победы следует отбросить принципы “политика - убийца разума” и “аргументы там - солдаты”
- На настоящей войне не стоит чураться использовать весь арсенал
- Оружие нейтрально само по себе, важно только если применяешь его ты или противник
- Нет идеального мира, которым управляет Законопослушный Нейтральный Космический Арбитр
- С принципом “я не хочу побеждать, если я неправ” невозможно удерживать власть
- LessWrong как пример, почему “левые” как движение - беззубые, бесхребетные, безвольные лузеры

Пересказ

I. Дискурс, основанный на швырянии грязью и лжи вредит обществу.

II. Призывы: «общественный договор» или «цивилизованных людей не прибегают к насилию». Аргументы: мы живем в более мирное время, конфликты не решаются еще одной мировой войной, мы достигли договоренностей (Женевская конвенция), влияние МЛК (ненасильственное сопротивление).

- Пример агрессивного ответа на насилие: Ирландская республиканская армия.

III. Томас Гоббс, Левиафан: о том, почему насилие - не единственный лучший способ решать проблемы. Принцип “око за око” при наличии суверена: чтобы не слить страну в унитаз, радикальные протестанты и католики договариваются не убивать друг друга, затем не клеветать друг на друга. Эквилибриумы без суверена: взаимный коммунитаризм (око за око) и божественная благодать (дискуссия без насилия, благородство среди врагов) - их комбинация создает социальные нормы (лгать - плохо).

- Левиафан - делегирование государства вопросу
- Фракция замечает что преследование правилам хорошо именно ей самой
- Божественная благодать: если вы что-то не понимаете, то вам нужно называть это термином, который бы напоминал вам о вашем непонимании, иначе вы будете думать, что всё объяснили, просто дав ему имя (отсылка к Юдковскому - сложность)

IV. Возделывание собственного сада. Проблема в “сорняках” внутри сада, а не в невозделанной почве за пределами. Проблема не сотрудничать в дилемме заключенного, призывая сотрудничать в дилемме заключенного.

V. Использовать ложь, оскорбления и харассмент по отношению к кому угодно нехорошо, даже если это поможет вам внедрять предпочтительные социальные нормы. Защита принципа отказа от насилия: примеры феминизма, трансгендеров, порнографии. Активная политическая позиция: я и все их новые либеральные друзья будут сражаться изо всех сил против тех, кто использует любые найденные сколь угодно малые различия между полами для того, чтобы бросить вызов основному либеральному убеждению о том, что человек имеет право на человеческое обращение к себе.

VI. Когда либералы вступают в ссоры, они всегда побеждают.

VII. Влияние Ганди (ненасильственное сопротивление).

VIII. Либерализм побеждает не огнём и мечом. Элуа — это бог цветов и свободной любви, и он страшен.

Вопросы

- Актуальны ли примеры благородства среди врагов на войне? Во второй мировой благородства уже не было, мы сейчас живем в другое время, чем как во времена Илиады или ПМВ

- Левые побеждали во всех войнах - примеры заканчиваются на ВМВ - игнорируется рост правого движения после этого
- Почему бы не убить видных расистских и сексистских политиков и интеллектуалов?

Ссылки

- Scott Alexander, In Favor of Niceness, Community, and Civilization, <https://slatestarcodex.com/2014/02/23/in-favor-of-niceness-community-and-civilization>
- Chris Hallquist, On some criticism of LessWrong, <https://web.archive.org/web/20160312160631/https://www.patheos.com/blogs/hallq/2014/02/on-some-criticism-of-lesswrong>.
- О пользе любезности, общин и цивилизации на русскоязычном форуме LessWrong, <https://lesswrong.ru/forum/index.php?topic=952.0>.

Обсуждение

- Игра про теорию игр: <https://notdotteam.github.io/trust>
- Пример про пустыню: https://lesswrong.ru/w/Критика_наивного_консеквенциализма
- Uncle Doc: <https://uncle-doc.livejournal.com>

Конспект обсуждения

О пользе любезности

ошибка выжившего: мы не знаем
примеры непобедившего
ненасильственного подхода

пример по man's land в ПМВ:
действия стали намного менее
бесчеловечные - давали время
растать раненым

платежная матрица: стратегии
кооперировать и враждовать -
дилемма заключенного

равновесие Нэша при первой
игре - враждовать

при итерации - становятся
выгодны новые стратегии

игра „Эволюция доверия“ про
теорию игр

Основные выводы:

- если ты и оппонент пользуетесь
разными эпистемологиями, даже
если вы сойдетесь в выводах - в
какой-то момент можете прийти
к разным выводам

- при наличии одинаковых
эпистемологий можно
договориться

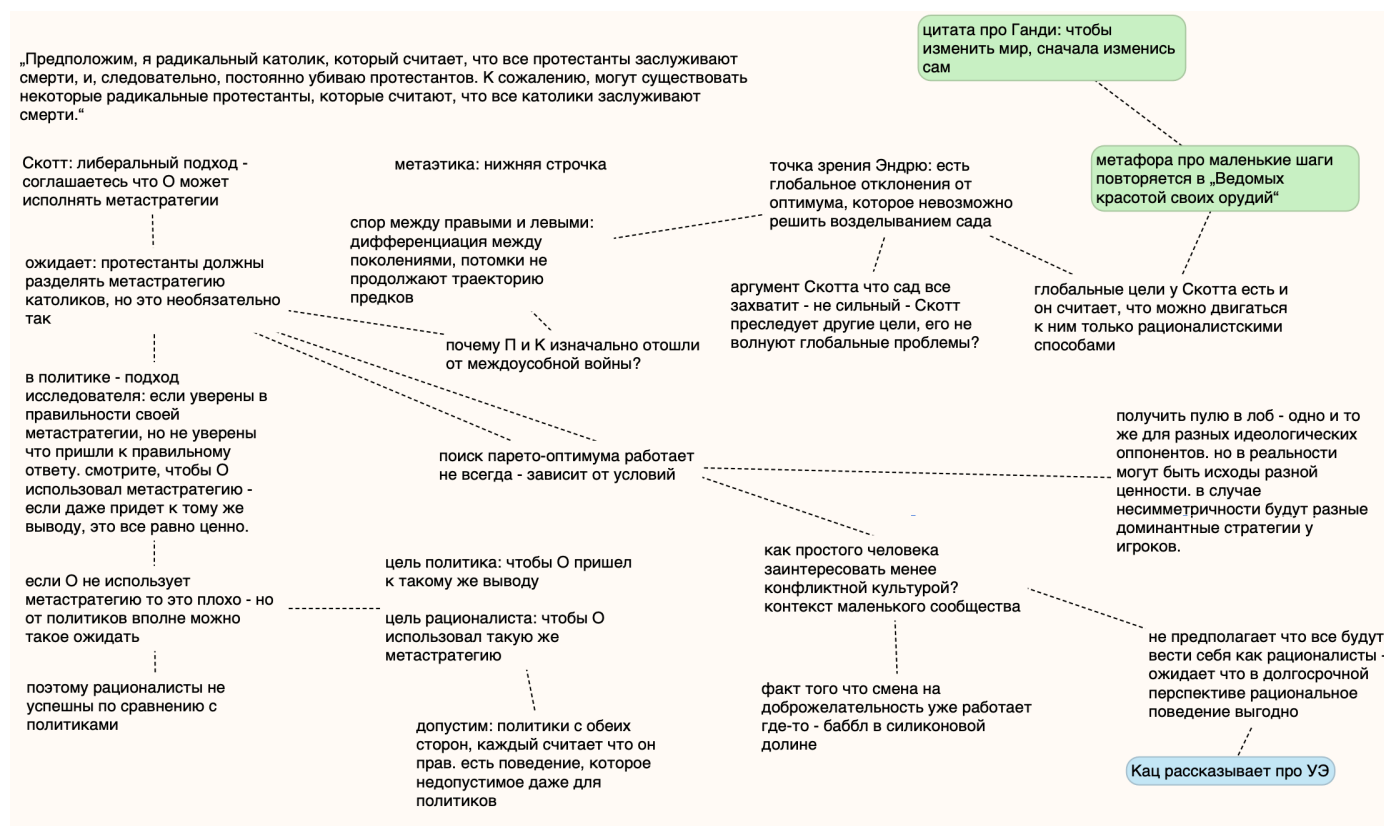
- что насчет средств
самозащиты? можно применять
насилие в таких экстренных
случаях

- дискуссия без насилия

вневременная теория принятия
решений: если мы определили
доминантную стратегию, стоит
ли применять ее всегда?

каузальная или наследственная
работает лучше в этом примере,
ВТПР используется для дилеммы
заключенного

пример в котором работает
ВТПР: про пустыню и проезжего



О рациональности. Скотт Александер, Прививка от сомнений. Элиезер Юдковский, Хранители Айн Рэнд (29.08.2020)

Почему эти эссе

Тезисы:

- Рациональность - про критический взгляд на собственные убеждения
- Рациональность - тоже система убеждений
- Безоговорочное восхваление рациональности пресекает критический взгляд

“Прививка от сомнений”

Основная идея: поиск нерациональности в гомеопатии и т.д. предотвращает поиск нерациональности в себе, прививает людей от сомнений в убеждениях - инокуляция установки. Если приводишь слабые аргументы как критику, оппонент их легко разрушает и укрепляется в своей уверенности.

1. Знакомый Скотта объясняет, почему важна рациональность - посмотрите на креационистов, гомеопатов, и верящих в лунный заговор. Скотт считает, что рациональность - про нахождение иррациональности в себе, а не в окружающих.
2. Разумное поведение в отношении креационистов/гомеопатов: не тратить время на анализ свидетельств "за", игнорировать представителей этих убеждений, не сомневаться в экспертном консенсусе "против".
3. Почему не стоит противопоставлять рациональность креационистам/гомеопатам:
 - В этих примерах легко определить иррациональность
 - Мы начинаем считать их репрезентативными примерами неправильных убеждений
 - Начинаем верить, что иррациональность очевидна
 - Затем начинаем верить, что неправильные убеждения не могут проникнуть в нас
 - Ошибочно полагаем, что игнорировать - верный подход к любому неправильному убеждению
 - Формируются вредные мыслительные привычки
4. Инокуляция: презентация слабых аргументов, напротив, укрепляет веру собеседника.
5. Не доверять экспертному консенсусу во всех случаях вряд ли комфортно. Но есть проблема в отсутствии воспитания критического мнения, когда все вокруг твердят обратное. Критики нужны для сдвига науки.
6. Настоящий кризис веры - не то же самое, что разоблачить креациониста/гомеопата.

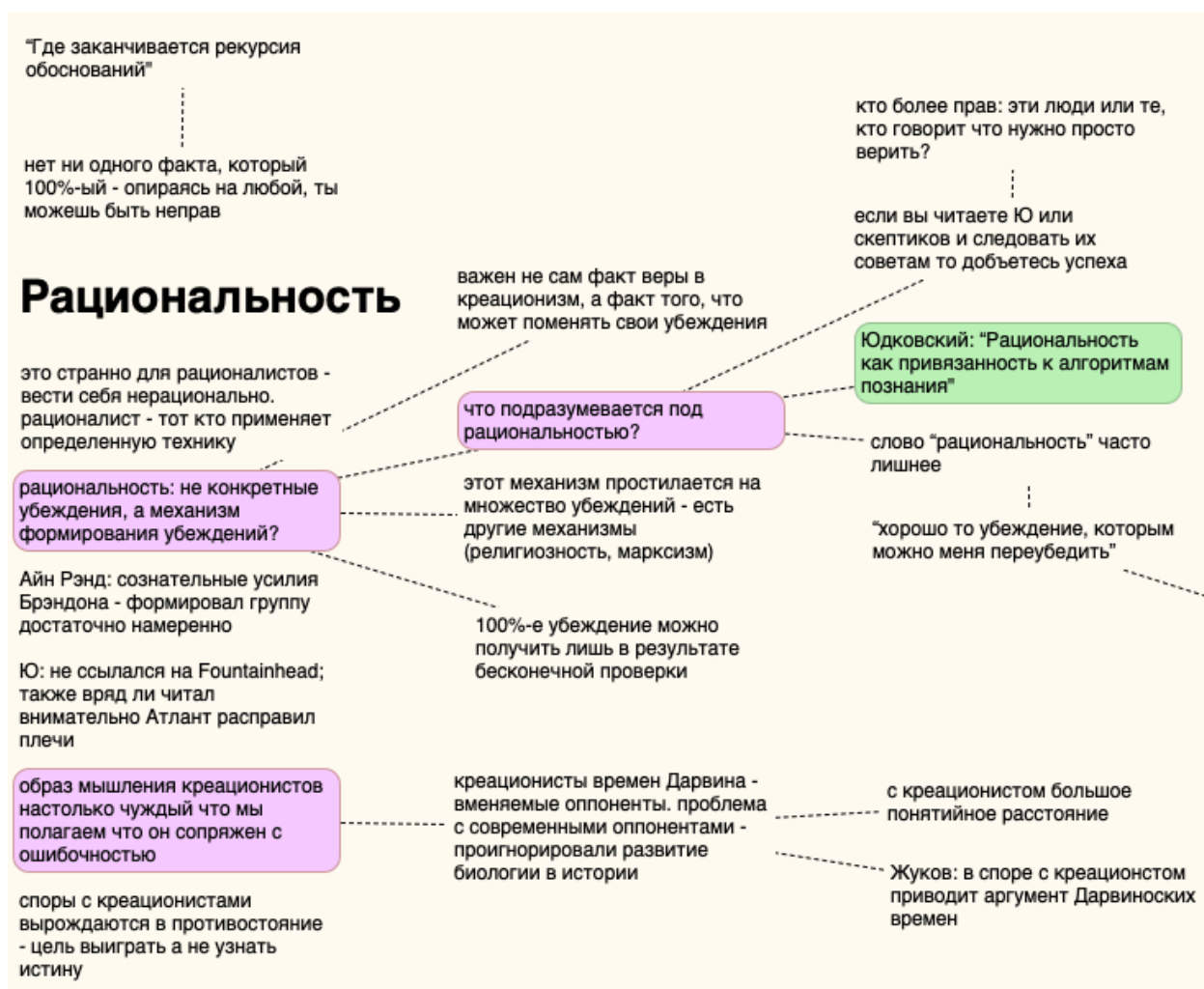
Отсылки

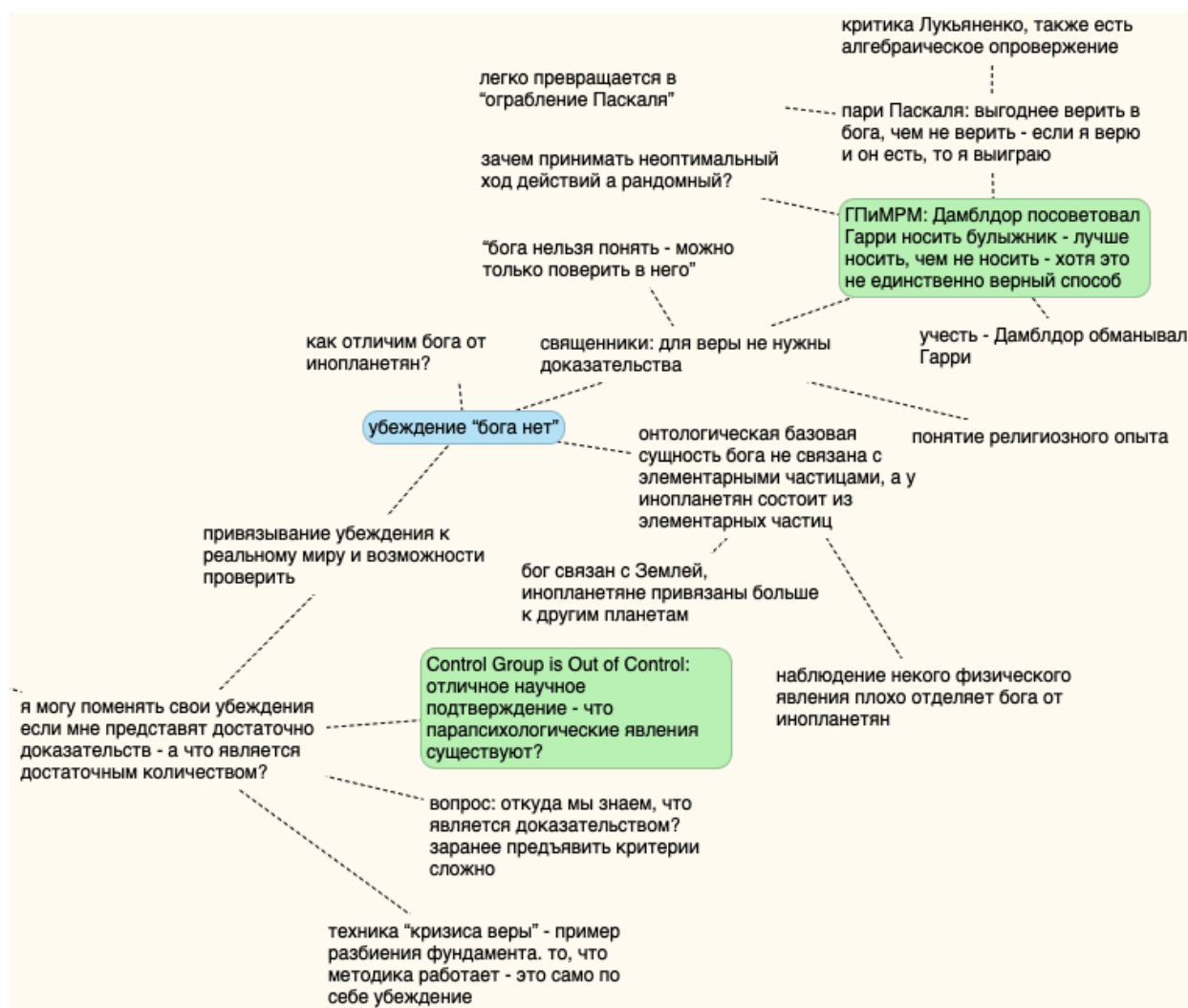
- [No, You Probably Don't Want 'Peer Reviewed Evidence for God'](#): публикации в научных журналах не имеют статус правды. Опасно основывать свою веру в отсутствие бога на научной проверке - если он появится и пройдет ее, тогда что?
- [Measles and the Inoculation Effect](#): описание процесса инокуляции.
 - Не хочешь, чтобы твои дети курили? Приведи им аргументы в пользу курения ("Это круто! Все это делают!") и спроси ребенка, почему они слабые. Когда ребенок столкнется с этими же аргументами от знакомых, ему не придется придумывать контраргументы на месте.
 - Обратная сторона медали: если сторонники твоего движения делают слабые аргументы в его защиту, когда ты придумаешь сильный аргумент, твои оппоненты уже перестанут вас слушать - они будут предполагать, что это очередной слабый тезис.
 - Если ты обвиняешь анти-вакцинщиков за вспышку кори, а затем выходит новость, что ты ошибался, и из-за этого больше людей решают изжить вакцины? Это будет плохо, не так ли?
- Также можно упомянуть отсылку к метаанализу парапсихологии из другого эссе Скотта: Control Group Is Out of Control

“Хранители Айн Рэнд”

1. Восхваление любого убеждения – даже рациональности – пресекает сомнения.
2. Культ вокруг философии объективизма Айн Рэнд – про него пишет Майкл Шермер – по мнению Юджовского ничем не особенный.
3. Убеждения давно умерших ученых устарели (Аристотель, Ньютон)
4. Рэнд восхищалась принципом необходимости превзойти старых мастеров, но не признавала никого, кто бы превосходил ее
5. Раскол объективистов после разделения Рэнд и Натаниэля Брэндена
6. «Изучайте науку, а не только меня!»
7. Прогресс нечестен! В этом вся суть!

Конспект обсуждения





Ведомые красотой своих орудий. Скотт Александер (16.08.2021)

Часть I

- Заметка начинается как комментарий к статье: <https://timharford.com/2017/03/the-problem-with-facts/>
- Обсуждается возникновение заблуждений (пример с голосованием 2016, пример с табачной кампанией)
- Статья цитирует исследование backfire effect, позднее поставленное под сомнение: <https://www.poynter.org/fact-checking/2016/fact-checking-doesnt-backfire-new-study-suggests/>
- Концепция "мудрых элит и серых масс".
- Посыл Скотта: следует сосредоточиться на передаче информации.

- Метафора класса, в котором каждый считает себя учителем.

Часть II

- Обсуждение статьи <https://www.currentaffairs.org/2017/03/debate-versus-persuasion> (о выборах 2016)
- Обсуждение тезиса "дискуссия с использованием чистой логики не работает"
- Тезис Скотта: дискуссии с использованием чистой логики вообще не было в обсуждаемом случае
- Список признаков внятной дискуссии:
 - Дискуссия должна быть диалогом встретившихся для неё людей (а не монологом с комментариями)
 - Оба участника должны быть, хотя бы в принципе, готовы поменять мнение
 - Диалог ведётся с соблюдением правил взаимной вежливости
 - Отсутствует элемент соревнования ("набора очков")
 - Дискуссия имеет внятную тему, и оба собеседника стараются не уходить от неё
- Наблюдение "люди игнорируют дискуссию без этих признаков" не поддерживает тезис "люди глухи к логике"
- Перемена мнения - очень постепенный процесс, не следует ожидать разворота на 180 за пятиминутный обмен
- Опыт Скотта дискуссии на его сайте (и ветке Reddit) на ту же самую тему
- Карикатурный пример аналогично невнятной психотерапии: "фактчек: пациент врёт"

Часть III

- Обсуждение проблемы: внятная дискуссия плохо масштабируется
- Цитируется исследование об обратном эффекте предъявления фактов - два исследователя получили разные результаты и провели новое совместное (adversarial) исследование
- Интересно что даже уверенность в собственной правоте не мешала бы такой кооперации
- Контраст: перепалка двух журналистов каждого со своих страниц
- "Не стоит ли перенаправить часть усилий с пропаганды правильной дискуссии на её проведение?"

Часть IV

- А с чего вообще стоит тратить силы на внятную дискуссию? Настаивать на своей правоте легко и приятно.
- Вводится идея "асимметричного оружия": такого, которое сильнее в руках хороших парней.

- Тезис: лёгкие и приятные меры одинаково сильны в руках хорошей и плохой стороны.
 - Фашисты дерутся так же больно, как и антифашисты;
 - Табачные компании снимают столь же эмоциональные фильмы, как и министерство здравоохранения
- Победа с использованием "симметричного оружия" по своей сути случайна (но если тебе повезло, не стоит смотреть в зубы дарёному коню)
- Пример с психотерапевтом и пациентом, который считает себя психотерапевтом, когда оба принимают таблетку.

Часть V

- Если вы верите, что факты не действуют на 50% людей, то откуда вы знаете, что вы в правильных 50%?
- Если у вас есть убеждение, что факты привели вас к нынешнему (правильному) мировоззрению, то вы считаете, что факты насколько-то работают
- Смена мнения - проявление истины - медленные процессы, которые к тому же зашумлены.
- Метафора сигнала и шума [ref: Нейт Сильвер]
- Вывод: для большей силы сигнала необходимо поднимать уровень здравомыслия (sanity waterline).

Обсуждение

Тезис: неравенство по знанию и пониманию фактов действительно существует

- Возражение: следует отличать "X знает больше" и "X говорит что знает больше"
- Встречно: поведение "что бы мне ни сказали, я буду делать как делал" существует
- В какой-то момент картина мира становится самостоятельной ценностью
- Уточнение: в случае конфронтации публичная приверженность своим утверждениям усиливается
- Ключевой момент: (идея) жонглирующей фактами элиты существует в головах
 - Курильщик может принимать сознательный обмен вида "риск рака за удовольствие"
- То, "действуют" ли факты сами по себе, зависит от ситуации в которой они предъявлены
- Разные люди по разному реагируют на факты в зависимости от ментальных привычек
 - Привычки, в свою очередь, могут существенно зависеть от (суб)культуры
 - Эффективность критикуемых в заметке приёмов, однако, от этого зависит меньше
 - Поэтому может быть полезно применять "убедительные истории" в заданной ситуации

Тезис: в политических разногласиях люди демонстрируют "ошибку сине-зелёного":

- Обобщаются черты фокус-группы "сторонников партии которая мне не нравится" на всех сторонников
- Соответственно, встречно, указание "существуют сторонники этой партии без таких характеристик" - трюизм
- Но - см. выше - есть более и менее умные/восприимчивые к фактам люди
- Указание: при разных терминальных ценностях выводятся разные "должен" из одного и того же "есть" (гильотина Юма)
- Предположение: разногласия во мнениях приводят к заниженной оценке интеллекта оппонента
 - Насколько вообще важно, что сторонники "синих" несколько более/менее умны?
 - Всеобщие дебаты вынуждены ориентироваться на уровень ниже среднего по обществу, а не на высший.

Тезис: множество способов поменять своё мнение шире класса "внятных дискуссий" (напр. чтение статей)

- Тезис2: если одна из сторон отказывается от диалога вообще, то все попытки ведения "внятной дискуссии" не приведут к смене мнения
- Риторика повышает готовность человека размышлять над темой (каковая готовность может быть очень низкой).
- Идея статьи: несмотря на существенно разный опыт и знания собеседников, они в некотором смысле "равноправны"
- Описанная в заметке проблема: люди выдают за "внятную дискуссию" то что ей заведомо не является
- Есть существенная разница между ситуацией "есть много людей без сильного мнения по предмету" и ситуацией "почти все уже определились и окопались в своих предпочтениях"

Если уверенность в своей правоте не препятствует кооперации, то что всё-таки препятствует?

- Ответ: отсутствие кооперации - "естественный" аттрактор; ему должно что-то противостоять ((суб)культура)
- Наблюдение: нет таких историй, где провал кооперации был бы удивителен
- Предположение: анти-кооперативные тенденции восходят к эволюционной психологии и менее выражены для маленьких и/или идеологически однородных групп
- Предположение2: нужно допустить что оппонент поменяет свою точку зрения, оно есть не всегда
- Дополнение: бывают очевидные теоретико-игровые аргументы против кооперации (напр. "дилемма заключённого")

- Если разные участники расходятся во мнениях о чём они, собственно, кооперируются, это тоже приводит к провалу
- Идея: кооперация как в описанном примере может вести к репутационным потерям для более уважаемой стороны
- Идея2: кооперация такого рода просто не пришла в голову

Восприятие: происходит подмена тезиса.

- Исходные статьи - "предъявление фактов укрепляет заблуждения", предлагается популяризация интереса к фактам
- Если считать "Научный метод от Сагана эквивалентен идее трансферинга реальности", то факты тоже симметричны
- Если факты работают скорее на "хорошую" сторону, то популярные статьи тоже работают на неё
- Ограничение: "Брекзит это плохо" - это не факт и не выводимое из (одних) фактов суждение.
- Ассиметрия - в отношении к тому что есть "факт", в методологии и отношении к дискуссии
- Расширение примера с пациентом: пусть таблетка разрушает определённую область мозга (тогда врач жертвует собой)

Восприятие: выводы статьи менее универсальны, чем, к примеру, выводы статей Юджовского

- Замечание: Скотта регулярно интересуют вопросы, близкие к уже рассмотренным в теории игр
- В частности: ситуация с дорогим выделением сигнала описана в модели "асимметричных стратегий"
- В реальном мире: люди могут стремиться защитить свою точку зрения (точку зрения своего флага), а не следовать за фактами. И могут стремиться не менять своё мнение.
- Восприятие: "кидать друг в друга фактами" - симметричное оружие, заметка пропагандирует отход от него.
- Замечание: к XX веку наука стала синонимом респектабельности в аргументации (но так было не всегда)
- Возражение: асимметрична не наука, а её плоды; частное свидетельство по встечавшимся окружениям: "учёные говорят" социально анти-аргумент

Нейт Соарес (06.12.2021)

Забота о большем, чем ты сам

Пересказ:

- Сменить чувство вины на чувство чего-то конкретного

- Заботиться о широком круге людей
- Зачем заботиться о широком круге?
- Большинство людей уже заботятся о широком круге
- Мы неидеальные эгоисты.
- Мы можем сопереживать ограниченно, поэтому мы думаем что забота то же ограничена.
- Забота это не про сопереживание а про правильные действия несмотря на отсутствие глубоких переживаний
- Откуда забота и чувства
- Чувства от первобытных инстинктов
- Забота от воспитания и эстетика - предпочтительно для Нейта
-
- Пример с нападением бомжа и собаки
- Чувства для собаки сильнее
- Стараться видеть в людях собак (с точки зрения заботы).

Обсуждение:

- легитимность обоснования подчинения эстетике всех остальных уровней. Терминальная ценность и все.
- Почему к собаке отношение более дружелюбное? С бомжом можно договориться а собаку убить придется
 - собака не понимает что делает.
 - без эмпатии одно чувство при виде собаки - Убей
 - первосистемное чувство почему они разные?
- Ценность консистентности в отношении чувств
- Мотивированная остановка
- попытка передать интуиции. консистентность превыше всего
- Чем больше понимания тем более человек похож на механизм, которому сложно сочувствовать

Непоследовательным быть можно

Пересказ:

- Излишне приверженным идее последовательности
- Боб не хочет помогать Алисе, но не знает почему.
- Справедливое наказание для разных людей
- Представьте себя арбитром внутри себя
- Если вы себя пересилили, то где-то потерпели неудачу
- Попытка быть последовательным, игнорируя части себя приведет к большим ошибкам

Обсуждение:

- Хейт против логики, т.к. она все время советует что, что внутри кажется не правильным
- Любительские судебные решения скорее приводят к проблемам

- Выученная эпистемическая беспомощность (Скотт Александер)
- логики бывают ненадежными
- Те у кого хороший когнитивный аппарат, у того хорошая логика
- При воспитании детей последовательности хороша. + желание сохранить авторитет и не признаваться в ошибках
 - Быть последовательным и проговаривать свои действия
- Почему Боб отказал?
 - он не знает
- Хочется не делать внутри противоборствующие стороны

Ещё не боги

Обсуждение:

- Чувство вины за то, вы занимаетесь чем-то хуже ваших представлений
 - Лишение агентности самих себя
 - Мальчикам преподают тяжелые упражнения - Ю.
 - Упражнение от Н С - это для пожилых
- Вина за то, что мозг не действует по их шаблону
- Обезьяна внутри нас старается, но у нее не всегда получается. Вам следует быть к ней добрее
 - Старается - не аргумент. Либо получается либо нет. Уровень ненависти к себе это важно
 - самоиндуцированная ненависть к себе или внешняя?
 - готовность отнести к себе пренебрежительно
 - Модель когда вообще никак не относишься к себе еще лучше
 - Замечание: человек с "героической ответственностью" без дополнительных костылей для бережного отношения к себе рискует впасть в апатию.
 - Возможно, цепочка replacing guilt, написана как результат того, что Нейт проникся добродетелью «воспринимать идеи серьезно» и серьезно воспринял идею героической ответственности. И написал цепочку о том, как смотреть на мир, чтобы не сломаться, глядя на него через призму героической ответственности и экзистенциальных рисков
 - вина - мешающая штука

Когда происходит выбор?

Пересказ:

- Возможности обезьяны ограничены, поэтому возможность не совсем соответствует реальному положению дел.
- Я мог бы остановиться и не играть. Мог должно возникать где нет заметного расходования силы воли.
- Представьте что у вас совсем нет силы воли

- Поставьте триггеры, например вы могли бы не запускать цивилизацию, или могли бы поставить будильник
- Научитесь выделять точки и смотрите где именно находится ваши "мог бы"

Обсуждение:

- мы имеет выбор когда мы агенты
- Мы снимаем ответственность с человека при физической невозможности совершить выбор

Лекция по книге “Рациональность: от ИИ до зомби” (28.02.2022)

- Мы неправильно мыслим
 - Когнитивные искажения
 - Якорение и корректировка
 - Прайминг
 - Контаминация
 - Фундаментальная ошибка атрибуции
 -
 - Нижняя строчка
- Как с этим бороться
 - Искать неверные убеждения
 - Искать убеждения основанные на слабых основаниях (заученные штуки из школы, которые мы не понимали)
 - Кризис веры
- О чем следует подумать - дополнительные трудности
 - Когнитивные искажения второго уровня
 - Понятийные расстояния
 - Дискуссии о политике
 - Рациональность второго уровня: самообман невозможен, когда ты узнал про искажение
 - Отсутствие систематического искусства
 - Как передать знание в прошлое
 - Избежать смертельные спирали
- Как передать знание - как поместить знание в голову ИИ
- Создание дружественного ИИ
 - Наивно ожидать что “придет ИИ, порядок наведет”