Машина у духові

Rationality: from AI to Zombies — Eliezer Yudkowski



Вперше опубліковано як записи у блогах Overcoming Bias та Less Wrong, з 2006-го по 2009-й рік.

Цей переклад — фанатський проєкт, що створюється з 2023-го року, з люб'язного дозволу авторів. Не для продажу. Не для комерційного використання. Всі права на оригінальний твір належать Еліезеру Юдковскі

Від перекладача — Данила Жирка



Красно дякую друзям зі спільноти pre post, які терпляче і вичерпно роз'яснювали складні моменти у книзі щоразу, коли я цього потребував.
Завдяки вам у перекладі куди менше помилок.

Від дизайнера — Дениса Лук'яненка



Долучитися до перекладу, вказати на помилку або просто подивитись, як все влаштовано можна на GitHub:
github.com/DanTheStrongworded/rationality-ua

Зверстано у Сторінкаторі 1.0.0



Гарнітури: EB Garamond, Geologica.
github.com/denlukia/storinkator

Версія перекладу 1.3 · 11 жовтня 2026

Зміст

7 Розум: вступ від Роба Бенсінґера

14 Проста математика еволюції
15 ІНТЕРЛЮДІЯ: Сила інтелекту
19 Чужий бог
27 Диво еволюції
32 В еволюцій мозок від курки, а витівки — на віки
37 Еволюції не стосуються ані корпорацій, ані нанопристроїв
43 Доеволюціонувати до вимирання
50 Трагедія групового добору
56 Критерій помилкової оптимізації
59 Виконавці адаптацій — не максимізатори пристосованості
63 Еволюційна психологія
68 Особливо витончений еволюційно-психологічний експеримент
73 Надстимули та занепад західної цивілізації
78 На уламках бога

84 Крихкі цілі
85 Віра в інтелект
88 Люди в кумедних костюмах
95 Оптимізація і вибух інтелекту
102 Духи у машині
107 Штучне додавання
114 Цінності термінальні та інструментальні
124 Нещільні узагальнення
128 Прихована складність бажань
136 Антропоморфний оптимізм
141 Втрачені цілі

148 Путівник словами для людей
149 Притча про кинджал
151 Притча про боли́голов
155 Слова — приховані умовиводи
158 Обсяг та зміст поняття
163 Скупчення подібностей
165 Типовість та асиметрична подібність
168 Структура скупчень у просторі можливих об'єктів
172 Замасковані запити
177 Нейронні категорії
183 Як алгоритм відчувається зсередини
188 Суперечка про визначення
194 Відчути значення
198 Аргумент від поширеного вжитку
203 Порожні позначення
206 Затабуюйте це слово
210 Замініть символ суттю
215 Помилки стиснення
219 Категоризація має наслідки
222 Приховані конотації
226 Аргумент «За визначенням»
230 Де ж провести межу?
234 Ентропія і короткі коди
238 Взаємна інформація та щільність у просторі об'єктів
245 Надекспоненційний простір понять та прості слова
252 Умовна незалежність та наївний баєсів метод
260 Слова — руків'я пензлів думки
263 Хиби прихованої змінної
266 Тридцять сім способів схибити зі словом

276 Вичерпний посібник з використання теореми Баєса

278 Геометричні візуалізації:
знайомство з баєсовою статистикою
284 Каскадні візуалізації та співвідношення шансів
291 Вступ до теореми Баєса: форма співвідношення шансів
299 Теорема Баєса: пропорційна форма
303 Надзвичайні твердження потребують
надзвичайно сильних свідчень
311 Звичайні твердження потребують свідчень звичайної сили
313 Теорема Баєса: логарифмічна форма шансів
322 Рухайтеся до найменш неочікуваної гіпотези
326 Теорема Баєса: векторна форма
332 Перегляд переконань як вилучення ймовірнісної маси
336 Теорема Баєса: форма ймовірностей
341 Баєсівський погляд на наукові чесноти

Розум: вступ від Роба Бенсінґера

Ви — це ваш розум. Такий порядок речей ставить людину у дивне становище.

У всесвіті не так багато розумів. І саме вам пощастило бути тією химерною штуковиною, здатною передбачати, планувати, зважувати і переглядати переконання, страждати, мріяти, милуватися сонечками або раптово захотіти манго. І навіть — виображувати власний розум *усередині свого розуму*. Можете аналізувати думки і старатися оптимізовувати їх під свої цілі.

Ви — розум, реалізований у людському мозку. А ще — попри всю його надзвичайну гнучкість, мозок піддається законам природи, моделюванню та рутині. Усе життя ваш розум слідує звичкам і навіть не помічає цього. Буває, звичкам з дивовижними наслідками. Тому коли модель розуму служить вам як слід, такий стан справ ми звемо «раціональністю».

Ви такі як є, генетично схильні проявляти певні форми раціональної та ірраціональної поведінки — усе завдяки вашим пращурам. Як і все живе на Землі, ви виникли з дуже давніх самовідтворюваних молекул. Початкові процеси відтворення були хаотичними й випадковими, але саме з них невдовзі виникли *відмінності* між молекулами-реплікаторами. І терміном «еволюція» ми описуємо зміни у цих відмінностях на відрізку часу.

Деякі з реплікаційних варіантів впливали на здатність до подальшого відтворення — це явище отримало назву «природний добір». І саме завдяки добору з'явилися організми, краще пристосовані до умов середовища наших предків. Те, ким ви є, сформувалося з відлуння випробувань та адаптацій, через які проходили покоління ваших предків.

Ось так і з'явилися ви — розум, висічений зі слабших умів, які прагнули розібратися у внутрішніх процесах, щоб краще поратися з ними. А для цього ви орієнтуєтеся на *свої* цілі, а не

вашого творця — еволюцію. Тож які практичні висновки та орієнтири ми можемо вивести з цього еволюційного підґрунтя?

Духи і машини

На поверхні, за своєю структурою та рушійністю, людський мозок нагадує звичайну механічну систему. Та коли заходить за розум, ми вже менш схильні уподібнювати його до об'єктів реальності чи органів тіла. Адже опорні категорії розуму — переконання, рішення, слово, ідея, почуття тощо — заледве схожі на категорії фізичного світу.

Прихопивши це спостереження, філософи минулого були носилися з думкою, що розум і мозок — засадничо відмінні явища. Філософ Ґілберт Райл назвав цей погляд «догмою духу у машині».[1] А втім, сучасні науковці та філософи, які відкинули дуалізм, так і не замінили його доладнішою передбачувальною моделлю мозку. У *практичному* сенсі наші цілі і бажання досі поводяться як відірвані від землі духи — наче та магістерія, відрізана від решти наукового знання. Ми можемо тут говорити про «раціональність», «біас» і «зміну думки», та якщо ці ідеї все ще не точні, не обмежені всеохопною теорією, наукова на позір мова не врятує нас від тих помилок, яких уже припускалися теоретики духів та сутностей.

Цікаво, що таємниці і містифікації довкола розуму затуманюють наше розуміння не тільки *людей* — вони кочують і до систем, які у галузях еволюційної біології та штучного інтелекту асоціюються з розумом та цілями. Якщо ми не можемо визначитися, з якого дзеркала нам дивитися на себе, певно, більше принагодитися можуть очевидно *не*людські процеси.

Повчитися можна у духів минулого, теперішнього і прийдешнього. Адже ці ілюзії — невигадані когнітивні події, які ми можемо вивчати і пояснювати. Якщо *здається*, що машиною кермує дух, це відчуття вже є результатом невидимої роботи машини.

1. Gilbert Ryle, *The Concept of Mind* (University of Chicago Press, 1949).

У «Простій математиці еволюції» — першому ланцюжку книги «Машина у духові» — ітиметься про розбіжність між нашою спадковою історією, сучасною біологією та людськими найкращими пориваннями. Ми підемо набагато далі, ніж заведено йти у вступах до еволюції для небіологів, що часто зупиняються на поверхневому описі природного добору.

У «Путівнику словами для людей» — третьому ланцюжку — обговоримо основний зв'язок між мисленням та формуванням понять. Опісля йде довша стаття-вступ до баєсового висновування.[2]

Між ними пролягає ланцюжок «Крихкі цілі»: він відходить від теми людського мислення та еволюції до ідеї розуму та цілеорієнтованих систем у їхньому найзагальнішому вигляді. Ці есеї служать другорядній меті — пояснити загальний підхід автора до філософії та науки раціональності, які значною мірою виходять з його досліджень штучного інтелекту.

Перебудувати інтелект

Еліезер Юдковскі — теоретик рішень і математик. Він займається засадничими питаннями загального штучного інтелекту і вивчає системи, які розв'язують загальні задачі. З першого есею «Бойове мистецтво раціональності», опублікованого у блозі Overcoming Bias, саме робота зі ШІ мотивувала Еліезера досліджувати психологію людської раціональності:

> «Моє нинішнє розуміння раціональності я здобув, коли намагався розв'язати проблему загального штучного інтелекту (ЗШІ). І задля успіху у цій справі раціональність потрібно опанувати настільки, щоб із самих лише зубочисток і канцелярських гумок зібрати повноцінного раціоналіста. Здебільшого проблема ШІ геть складніша за

2. У перекладі замість есею «Інтуїтивне пояснення теореми Баєса (https://www.lesswrong.com/s/SGB7Y5WERh4skwtnb/p/XTXWPQSEgoMkAupKt)» поданий розділ «Вичерпний посібник з використання теореми Баєса» з дванадцяти есеїв (автор вважає його кращою версією згаданого есею). — Прим. пер.

особисте мистецтво раціональності, проте де в чому він простіший. Що стосується бойового мистецтва розуму, наше завдання — набути в реальному часі методичної навички смикати потрібні важелі у потрібний момент всередині великої мисли́вої машини з уже заданими внутрішніми параметрами, вносити зміни до яких кінцевий споживач не має змоги. Деякі з цих машин налаштовані під тиском еволюційного добору таки́м чином, що прямо суперечить нашим заявленим цілям використання. Наш свідомий вибір — пошук істини, хоч мозок людини і має вбудовану програму для обґрунтування брехні. [...]

Спроба зібрати особисте мистецтво раціональності з науки раціональності може виявитися недолугою: уявіть-но, як хтось витворює бойове мистецтво на основі абстрактної теорії фізики, теорії ігор та анатомії людини. Але люди здатні до рефле́ксій. Нам таки притаманний інстинкт самоаналізу. Внутрішнє око видю́ще, хоч і бачить воно невиразно, із систематичними спотвореннями. Саме тому, щоб коригувати наші ментальні рухи і посилювати метакогнітивні навички, інтуїції не обійтися без науки і теоретичних знань. Ми не пишемо комп'ютерну програму для того, щоб примусити маріонетку до бойового мистецтва. Ми маємо навчитися керувати власними ментальними кінцівками, тому слід пов'язати теорію з практикою. Ми повинні зрозуміти, яке значення має наука раціональності особисто для нас, для осібного повсякдення».

Наскільки я розумію, з погляду Юдковскі, годі розповідати про людську раціональність без цікавих вкраплень про ШІ, причому це працює в обидві сторони.

Юдковскі передбачає, що штучний інтелект врешті-решт перевершить людей у процесі «вибуху інтелекту» (так називають сценарій, за якого самозмінюваний ШІ, перепрограмовуючи самого себе, покращує власні здібності, чим запускає стрімку

послідовність подальших самовдосконалень). Іноді замість «вибуху інтелекту» послуговувалися поняттям «технологічна сингулярність» — аж до січня 2013 року, коли Науково-дослідний інститут машинного інтелекту (MIRI) назвали «Інститутом сингулярності». Він якраз проводив річну зустріч щодо сингулярності. Відтоді Юдковскі послуговується старим поняттям Ірвіна Ґуда «вибух інтелекту», яке виділяє його погляди від решти футуристичних передбачень, наприклад, тези Рея Курцвáйля про експоненційний технологічний прогрес.[3]

Виглядає, що такі технології, як надлюдський інтелект, можуть призвести до масштабних потрясінь у суспільстві як у кращий, так і в гірший бік. Тому Юдковскі почав говорити про «теорію дружнього ШІ», в якій посилається на техніки вирівнювання цілей ШІ з цілями людськими. Та наразі дуже мало відомо про те, коли може з'явитися програмне забезпечення ЗШІ і які заходи безпеки будуть дієвими. Не випадає бути цілком певним, чи сучасні автономні системи ШІ працюють коректно і годяться для практичного використання. Відповідно, і сучасні заходи заледве можна узагальнити до розумніших та адаптивніших систем. Тому концепція «дружнього ШІ» ще не набула вигляду чітко визначеного набору цілей і складається радше з математичних та філософських питань.

Станом на 2015 рік погляди Юдковскі на майбутнє ШІ досі обговорюються у колі футурологів та галузевих й університетських дослідників, серед яких ще не сформувався консенсус. Прикладом, у своїй книзі «Суперінтелект» Нік Бóстром викладає широкий огляд багатьох моральних і стратегічних питань, які почнуть поставати з появою надлюдського інтелекту.[4]

---

3. Irving John Good, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, ed. Franz L. Alt and Morris Rubinoff, vol. 6 (New York: Academic Press, 1965), 31–88, doi:10.1016/S0065-2458(08)60418-0.

4. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

В одному з розділів книги «Штучний інтелект: сучасний підхід»[5] Стюарт Ра́ссел та Пітер Но́рвіґ наголошують на тому, наскільки технологічно складно визначити хорошу поведінку для високоадаптивного ШІ:

> «[Юдковскі] стверджує, що дружність (відсутність бажання завдавати людям шкоди) слід запрограмувати від початку, але ж розробники мають враховувати, що їхні задуми не досконалі, а роботи ще вчитимуться і розвиватимуться. Розробити для змінюваних систем ШІ механізм стримувань і противаг — завдання надзвичайно складне, адже попри зміни він має зберігати дружність. Стала функція корисності не варіант, оскільки обставини і наші очікувані відповіді на ці обставини теж змінюються з часом».

Своєю чергою, стурбовані загрозливим для людства прогресом у сфері ШІ, нано- і біотехнологій, Нік Бостром та Мілáн Чíрковіч склали першу академічну антологію «Ризики глобальних катастроф».[6] З них найбільш екстремальні ризики, *екзистенційні*, можуть вилитися у нескінченний застій чи навіть вимирання людства.[7]

Що люди загалом, що експерти, як правило, *надзвичайно кепсько* прогнозують важливі події майбутнього (і поява технологій не виняток). У розмовах про раціональність одна з цілей Юдковскі — з'ясувати, котрі біаси перешкоджають нам як прогнозувати значні зрушення, так і готуватися до них. До то́му «Ризики глобальних катастроф» Юдковскі долучився статтями «Потенційний негативний вплив когнітивних біасів на

5. Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010).

6. Bostrom and Cirković, *Global Catastrophic Risks*.

7. Приклад такого гіпотетичного екзистенційного ризику — «сірий слиз»: за цим сценарієм молекулярні роботи, здатні до самореплікації, витіснивши всі живі організми, повністю поглинають Землю. — Прим. авт.

судження про глобальні ризики[8]» та «Штучний інтелект як позитивний і негативний фактор глобальних ризиків[9]», які постали з його досліджень у когнітивних науках та ШІ. У Кембриджському посібнику зі штучного інтелекту Юдковскі та Бостром долучилися розділом про близькі та далекосяжні виклики під назвою «Етика штучного інтелекту».[10]

Хоча у книзі йдеться про *людську* раціональність, тема ШІ слугує простою ілюстрацією аспектів людського мислення. Прогнозування появи далеких у часі технологій — одне з важливіших застосувань баєсової раціональності, за допомогою якої можна змоделювати коректне міркування навіть за умови браку чи двозначних даних.

За творінням можна багато чого сказати про творця, а знайомство з творцем дає ключі до розуміння творіння.

Тож і почнемо з того, що наш творець може розказати про дизайн наших розумів.

8. https://intelligence.org/files/CognitiveBiases.pdf

9. https://intelligence.org/files/AIPosNegFactor.pdf

10. Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William Ramsey (New York: Cambridge University Press, 2014).

Проста математика еволюції

ІНТЕРЛЮДІЯ: Сила інтелекту

Щодня наші черепи перевозять сірувату глевку тканину вагою близько 1,3 кілограма, схожу на зібганий туалетний папір.

Бридка грудка менш за все нагадує наймогутнішу матерію у відомому всесвіті. Якщо ви в очі не бачили підручник з анатомії, мозок на тротуарі ви, скоріше, назвали б «фубльою», а потім перевірили підошву. Арістотель той узагалі думав, що мозок охолоджує кров. А все тому́, що та грудка не *виглядає* загрозливою.

П'ять мільйонів років тому день належав предкам левів, ніч — далеким родичам вовків. Альфа-хижаків еволюція озброїла зубами і пазурами — ріжучими лезами з міцними м'язами. На свій захист їхні жертви обросли грубими панцирами, гострими рогами, токсичною отрутою і маскуванням. Ця війна тривала сотні епох, одні види зброї мінялися іншими. Багато хто не пережив цю війну, однак досі не було жодного натяку на переможця. Одні види ховалися у панцир — інші навчилися дробити їх; на отруту одних інші виробили протиотруту. Кожен вид займав свою екологічну нішу. Хто взагалі зміг би жити у морі, небі й на суші водночас? Ніякої передумови для виникнення суперзброї, і відповіді на неї — суперзахисту. Ніхто й подумати не міг, що хтось на таке спроможний.

А потім прийшов День глевких грудок.

Ні, у них не було броні, пазурів чи отрути.

А якби вам показали відео випробування ядерної бомби й сказали, що її створила земна форма життя, ні за що на світі ви б не подумали, що ядерний гриб — творіння отієї глевкої субстанції. Зрештою, радіацію вона не випромінює.

З самого початку глевкі грудки не мали ні винищувачів, ні кулеметів, ні рушниць, ні мечів. Жодної вам бронзи чи заліза. Ні молотків, ні ковадел, ні обценьків, ні ку́зень, ні копалень. Були лише слабенькі пальчики, якими навіть дерево не зрубати — не те що гору звернути. Кому взагалі може нашкодити

той глевкий слимак? От камінь — його обробляють залізними те́слами. А ті грудки не виділяють залізо. У дикій природі годі було знайти залізні леза з руків'ями для немічних пальчиків. І температурою своїх тіл розплавити метал вони теж ніяк не могли. Та що тут думати!

А ДНК? Як своїми товстими пальцями вони примудрилися змінювати генетичну інформацію? Ще скажіть, що вони атом гідрогену тими пальцями можуть узяти. *Технічно* що глевкі грудки, що ДНК — це все один світ, що керується єдиними законами фізики, однією великою павутиною причиновості. Як накажете перестрибнути з одного рівня на інший, га?

Навіть якби глевкі штуковини *одного дня* навчилися чогось подібного, на це пішли б тисячі тисячоліть. Повірте: на відтинку розвитку Життя один рік не прирівняти навіть до одного цокання годинника. Безумовно, *технічно* рік дорівнює шестистам октильйонам[1] секунд (за планківським часом). Але у світі нічого не може відбутися менш ніж за шістсот октильйонів планківських секунд, тому це спірне питання. Коли глевка грудка вміє тільки бігати по савані, у неї ще десять мільйонів років не з'явиться жодних шансів навчитися літати через континенти; а там уже *ніхто* не зміг би розмножуватися із такою постійністю.

Тепер поясніть мені, чому штучний інтелект, не маючи фізичного тіла, не в змозі вдатися до якогось штука́рства в інтернеті.

Адже — уже не раз помічав — асоціація зі словом «інтелект» — думка, що виникає на першій секунді після почутого слова, — часто формує інстинктивну реакцію на вибух інтелекту. Здебільшого люди знаходять ключове слово «інтелект» і витягають образ *стереотипного розумника* — ґросмейстера, який ніяк не наважиться запросити дівчину на побачення, чи професора, який не виживе за межами академії.

«Успіх у кар'єрі не одним інтелектом здобувається», — так кажуть, ніби харизма секретується з нирок. «Інтелект не

1. У році $3{,}15\times10^{7}$ секунд. Планківський час дорівнює $5{,}39\times10^{-44}$. Відтак планківський рік дорівнює шестистам октильйонам (число з сорока вісьмома нулями) секунд. — Прим. пер.

зрівняється зі зброєю», — наче автомати ростуть на деревах. «Де ж той штучний інтелект братиме гроші?» — так питають, ніби перші *homo sapiens* збирали доларові купюри після грошового дощу й обмінювали їх у лісових крамничках. Ринкова економіка прийшла не з дикої природи. Бджоли не продають мед за переказ на картку. Гроші *придумали* люди, і вони досі існують — не для мишей чи ос, а для *нас* — бо ми в них віримо.

Досі мені доводиться пояснювати, що прообраз інтелекту — не Да́стін Хо́ффман з фільму «Людина дощу». Людина. Інтелект уособлює людина. Ота глевка речовина, яка може миттєво вибухнути у вакуумі, залишає сліди на поверхні Місяця. Всередині неї ховається сила шукати і знаходити, здавалося б, немислимі шляхи у великій павутині причиновості, сила, що зветься креативністю.

Люди — особливо ризикові інвестори — знай питають, який їм буде *зиск* з того ШІ, якщо інституту таки вдасться його розробити. Не ті питання турбують голови інвесторів.

Може, варто взагалі дивитися в іншу сторону. Адже якщо додати трохи креативу, то людина ось уже планує політ на Місяць, думає про ліки проти віспи, розробляє комп'ютер. *Усе це* здається практично неможливим, хай би скільки сантиметрів від ваших очей мешкала та субстанція. А все тому, що глевка грудка так і не збагнула глевку грудку.

Люди не бачать достоту, як працює та машинерія; сила інтелекту не очевидна, а від того виобразити її важче, ніж вогняну вежу, що відправляє корабель на Марс. Сама лиш думка про відвідини Марсу захоплює уяву. А от якщо хтось пообіцяє можливість дістатися Марсу, а ще — загальну теорію фізики, доведення гіпотези Рі́мана, ліки проти ожиріння, раку, старіння та дурості, нічого, крім поблажливих поглядів, він не отримає.

Звісно, не без скепсису, але проблема в тому, що інтелект годиться лише для небагатьох речей, походить з обмежень уяви. Чи міг хтось подумати собі, до чого одного дня дійде розум? Можливо, ми навіть не *знаємо* своїх справжніх проблем.

А втім, побачити, куди може завести цей процес, — завдання не з легких, як і уявити, що заразо́м хтось знайде рішення для таких прозаїчних проблем, як ожиріння, рак та старіння.

Аж тут глядіть: ло́вка грудка і ліки проти віспи знайшла, і літаки сконструювала, і пшеницю виростила, і вогонь приборкала. Хоч нинішні науковці й досі не можуть узгодити, як саме функціонує та ловка грудка, від того вона не перестає робити своє, грудкове. Нерозуміння феномена — це подія, що свідчить про стан свідомості, а не про власне феномен. З порожньої мапи не слідує порожня територія. Незалежно від того, чи хтось розуміє, яка сила залишає сліди на Місяці, сліди від того нікуди не зникають. Сліди видимі на видимому Місяці й залишені невигаданою силою. Якби ж то ми могли збагнути її досить глибоко, що б завадило нам створити ту силу? Інтелект не менш реальний, ніж струм. Проте набагато потужніший, небезпечніший, залишає набагато глибші сліди у ще не дописаній історії всесвіту. І зрозуміти, як відтворити його, буде все ж трішки складніше, ніж сконструювати електрогенератор.

Чужий бог

«У теорії еволюції цікаве те, — писав Жак Моно́, — що кожен думає, що розуміє її».[1]

При погляді на природу людина в усьому вбачає *провидіння.* Лапи дані кроликам для бігу, щелепи лисицям — для відгризання плоті. От тільки сови — і не тільки вони — насправді не ті, чим здаються...

Задовго до Дарвіна причина усього цього *задуму* була величезною загадкою для науки. Богомільні казали: «Це створив Бог» і відразу нараховували собі п'ятдесят бонусних балів за згадку (хоча, може, я чогось не розумію). До Дарвіна ця гіпотеза видавалася притомною. Про це ж і говорив Вільям Пе́йлі: з того, що ви знайшли годинник у пустелі, можна виснувати існування годинникаря.[2]

Але потім приходить розуміння: коли замість того, щоб відбирати зразки, ви хапаєтеся за *очевидне провидіння,* воно вже не вписується у юдео-християнську картину з одним милосердним Богом. І виявляється, що лисиці створені саме для того, щоб полювати на кроликів, а кролики — тікати від лисиць. Яка вавка в голові спонукала Бога до такого інженерського рішення?

Розробляючи духовку-тостер, мені в голову навіть не прийде сконструювати її так, щоб одна частина доставляла електроенергію до котушок, а друга перешкоджала першій. Навіщо взагалі так робити? *Що за істота* створює екосистему, в якій живуть хижаки і здобич, віруси і бактерії? Та навіть кактус, призначений нібито для напування пустельної фауни, навіщось обростає тими жахливими колючками.

А може, така екосистема була би більш зрозумілою, якби її створив не єдиний *Хтось*, а пантеон богів (скажімо, індуїстських

1. Jacques Monod, On the Molecular Theory of Evolution (1974), reprinted in Mark Ridley (editor) Evolution (1997) p. 389.

2. 1802 book Natural Theology or Evidences of the Existence and Attributes of the Deity.

чи синтоїстських). Так принаймні зручніше пояснити всюдисуще провидіння і пов'язані з ним проблеми: світ постав на перетині цілей різних богів. Кролика і лисицю, наприклад, створили боги, які конкурували між собою. Цікаво, чи хтось використовував такий, здавалося б, прекрасний аргумент на користь індуїзму проти християнства. Певно, що ні.

Те саме із заявленою (і тільки заявленою) милосердністю юдео-християнського Бога. Адже увесь той закладений *задум* виглядає відверто жорстоким. Вивчаючи їздців-іхневмонів (*Ichneumonidae*), котрі жалами паралізують жертву, щоб живцем згодувати її своїм личинкам, Дарвін уже підозрював, що Творець не вписується у шаблон: «Не можу повірити, — писав Дарвін, — що милосердний і всемогутній Бог заклав ув іхневмонів явний намір паразитувати на живих гусеницях аж до їхньої смерті або що коти мають ганятися за мишами».[3] Цікаво, чи хтось раніше згадував цей прекрасний аргумент на користь манихейства проти монотеїзму.

Але тепер усе змінилося: достатньо сказати «еволюція».

Боюся, саме так дехто і засвоює отаке «наукове» пояснення: еволюція перетворилася на магічну фабрику з виробництва провидіння у Природі. У першій книзі я вже писав про Шторм, яка через *одну мутацію* набула сили метати блискавки. Як таке сталося? Та от, бачте, є така собі «еволюція», яка бозна-як вкачує провидіння у Природу, і якісь «мутації» призводять до таких змін. Тож якщо гени Шторм мутують *аж до такої міри*, вона зможе метати блискавки. Радіація — ще одна популярна причина супермутацій: що сильніше опромінення, то більші мутації. Логічно ж.

Та насправді еволюція забороняє *будь-яке* провидіння у Природі. І саме через відсутність вседозволеності еволюція придатна до емпіричного тестування. Якщо еволюційною біологією можна пояснити духовку-тостер (а не, скажімо, дерево), від неї немає користі. Суть теорії еволюції зовсім не в тому, щоб

3. The Owner of All Infernal Names: A treatise on the existence of our Omnimalevolent Creator.

показувати на Природу і говорити: «Ось тут задум такий-то» або «Так забаглося еволюції!». Адже сила теорії в тому, що вона забороняє: якщо ваше пояснення годиться для будь-якого наслідку, у вас насправді ілюзія знання.

«Багато небіологів, — помітив Джордж Вільямс, — думають, що тріскачки ростуть на хвостах гримучників заради їхнього ж блага».[4] Бррр! *Такий* задум не годиться. Еволюція не перетворює один вид на користь випадкового реципієнта — такі спалахи провидіння просто не передбачені.

Еволюція починається зі систематичної кореляції між різними способами, якими гени будують організм, та кількістю реплікацій тих генів, які передаються наступному поколінню. Щоб на хвості гримучників почали рости тріскачки, у кожному наступному поколінні має збільшуватися кількість відповідних генів (точніше — генів для поступово складніших тріскачок; якщо я почну описувати усі тонкощі еволюційної біології, нам і дня буде *мало* усе обговорити).

Годі шукати фею Еволюцію, яка продивляється стан екосистем, а потім їй приходить у голову підвищити частотність генів, які формують тріскачки для гримучих змій.

Підозрюю: саме тут шпортаються багато людей. Вони знають, що «корисні» гени поширюються частіше, але в них ця «корисність» витікає у провидіння. Хоч вони й не мають на увазі фею Еволюцію, все ж питають, котрі гени виявляться «корисними», ніби ген гримучника був би «помічним» ще комусь, крім нього самого.

Головне — феї Еволюції чи будь-якої іншої зовнішньої сили, що *визначала* б поширення генів, *не існує, позаяк* усі зміни ініціюють самі гени.

Гени, завдяки яким (поступово) розвинулися тріскачки, поширилися саме *через наявність* тих тріскачок у гримучих змій. Це припущення *імовірне* через високу виживаність гримучників з тріскачками порівняно зі зміями, які частіше спарювалися, чи

4. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.

мали братів, які краще розмножувалися, чи якимись іншими рисами.

Можливо, хижаки остерігаються звуку, що видає тріскачка, і не наступають на змій. А може, тріскачка відвертає увагу від голови змії. (Згадати тільки слова Джорджа Вільямса: «Результат сутички між собакою і гадюкою більшою мірою залежить від того, чи собака вхопить змію за хвіст чи голову».[5])

Окрім зміїної тріскачки, існують і вигадливіші способи реплікації і передачі генів наступному поколінню. От у ваших братів і сестер половина ваших генів. Ген, що спричиняє витрату однієї одиниці ресурсу на користь брата, завдяки чому зберігаються три одиниці його спільних копій, може підвищувати власну репродуктивну успішність, навіть якщо це відбувається ціною організму-носія. (Розлогіше тільки у книзі про еволюційну біологію.)

Основне, *результат дії гена* повинен *призводити* до поширення цього гена у наступному поколінні. Не фея Еволюція приносить його ззовні. Ніхто не вирішує, який з генів «корисний» чи може поширюватися частіше. Маємо справу лише з причиною й наслідком, що починаються з самих генів.

Звідси й оці дивні конфлікти у провидінні та часту жорстокість у Природі. Не потрібен навіть пантеон синтоїстських божеств.

Чому ж так багато видів воюють один з одним за своє місце у Природі? Бо, знову ж, годі шукати однієї еволюції, яка керує усім процесом. Тих «еволюцій» стільки, скільки репродуктивних популяцій. Кролячі гени поширюються серед кроликів, лисячі — серед лисиць. І лисячі гени, що створюють лисиць такими, щоб вони ловили кроликів, передаються наступному поколінню. Кролячі гени, які створюють кроликів, швидших за лисиць, *природним шляхом* передаються наступному поколінню кроликів. Ось так еволюціоністи і дійшли до терміна «природний добір».

5. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.

Чому Природа така жорстока? Спостерігаючи за іхневмонами, людина прийде до висновку, що зжерти здобич живцем — це жорстоко. Якби вам довелося піти на таке, ви могли хоча б позбавити її страждань. Хіба тій осі щось коштує принаймні паралізувати свою здобич? А як щодо старих слонів, які помирають з голоду, коли у них випадають останні зуби? Ті слони вже не будуть розмножуватися. Що ж коштує тій еволюції — точніше еволюції слонів — вбивати їх одразу, а не мучити довгими стражданнями? Чому еволюція не може виробити якусь анестезію для слонів чи хоча б стимулювати їхні нейрони так, щоб перед смертю вони бачили приємні сни? Нічого. Що так, що так слони від того більше не розмножуються.

Якби ви пробували розв'язати конфлікт інтересів зі своїм ближнім, у вас була б хороша позиція для перемовин — переконати його не склало б труднощів. Це коштувало б так мало — знеболити жертву, дати слону померти без муки! Будь ласка, молю-прошу: зробіть... гм...

Вам нема з ким сперечатися.

Уже не вперше люди набувають бажання одним способом, а вже потім вигадують зовсім інший спосіб, щоб пояснити або виправдати собі цей вибір. Вони думають так, ніби існує фея Слонова еволюція, яка у постійному пошуці (а) найпридатніших рис для слонів, (б) способів *обґрунтувати* ті риси перед еволюційним Ревізором, який (в) не зацікавлений у падінні фертильності, але (г) не проти ідеї усунути біль зі смерті, щоб не нашкодити генам.

Ніхто не зацікавлений у такому.

Люди ж, які часто переймаються добробутом тварин, раді аргументувати, що різні прояви доброти аж ніяк не шкодять фертильності. На жаль, слонова еволюція не послуговується цим алгоритмом; вона не добирає хороші гени, *якими можна переконливо обґрунтувати* вищий рівень фертильності. Просто гени, що частіше реплікуються, збільшують імовірність своєї появи у наступному поколінні. Вона — що товща води, яка стікає зі скелі.

Людина, дивлячись на Природу, береться собі думати про всі ті способи, якими *ми* спроєктували б організми. Тоді ми починаємо придумувати виправдання, чому наші покращення у дизайні підвищили б фертильність. Такий собі політичний інстинкт, в якому закладена спроба продати свій найкращий спосіб, що відповідає вимогам керівництва.

Після чого такі еволюціоністи-любителі висувають щонайрозмаїтіші дивовижні і *цілковито безглузді* передбачення. Тому що ті любителі креслять свою нижню риску — що важливіше, ставлять свої передбачення у простір гіпотез — якимось *своїм* методом, відмінним від того, яким це роблять еволюції.

Якби смакові рецептори проєктував інженер, він встановив би лічильник наявних і потрібних організму поживних речовин. Наприклад, якщо рівень жиру на мінімумі, мигдаль або чизбургер повинен мати неперевершений смак. А якщо ви починаєте набирати вагу і вам бракує вітамінів, вам більше смакує листя салату. На жаль, *феї Людської еволюції,* яка б завчасно все спланувала і розробила загальну систему для будь-яких надзвичайних випадків, просто не існує. У середовищі, де жили предки людей, постійно бракувало калорій. Тому гени, носії яких любили калорії, поширилися так легко.

У нас утілена історія організмів, які *власне* вижили і розмножилися, а не тих, яким *слід було* вижити чи розмножитися.

Сітківка ока людини має зворотне розташування: світлочутливі клітини звернені від рогівки, а нервові волокна прямують від зовнішнього шару до внутрішнього, проходячи крізь сітківку до мозку. Звідси й сліпа пляма. При погляді на сітківку інженер хапається за голову, адже в інших організмах сітківка розвинулася незалежно і розташована правильною стороною. Чому б не переробити ту сітківку?

На жаль, *жодна* мутація не переробить *цілу* сітківку одним махом, на відміну від інженера, який може переробити декілька частин водночас або запланувати майбутні коригування. Та якщо бодай одна мутація ламає життєво необхідну частину

організму, байдуже, наскільки дивовижну надбудову може створити фея — організм помре і більше не передасть гени.

Якщо розвернути сітківку і не переробити напрям нервових з'єднань і хід зорового нерва, вся система перестане функціонувати. Байдуже, що це лиш один крок у перероблюванні сітківки. Організм утратив зір. Еволюція не вирізняється передбачливістю — вона лише зберігає гени організмів, яким *уже вдалося* лишити потомство. Вона така ж сліпа, як і наполовину перероблена сітківка.

Пам'ятаєте, що казав Вільям Пейлі, що з годинникового механізму випливає існування годинникаря? Були і такі, хто думав, що життя «виникає саме собою», без потреби в оптимізації, — що миша може постати просто з купи соломи і брудної білизни.

Якщо спитати, хто мав *більшу* рацію — теологи, які відстоювали ідею Бога-Творця, чи інтелектуально обмежені атеїсти, які вірили у з повітря створених мишей, — то перемогу слід віддати теологам: еволюція, хай і не бог, але значно ближча до нього, ніж до чистого випадку. Мутації випадкові, зате добір — *закономірний* (і ні, тут усе ще немає жодної розумної феї, що перебирає гени пальчиками). Отже, маємо ненульову кореляцію між генами і частотою розмноження організмів. Ось уже кілька мільйонів років ця кореляція в сумі дає щось надзвичайне. Її хоч і не назвеш богом, але вона ближча до нього, ніж до білого шуму.

Багато в чому еволюція подібна до теології. «Боги онтологічно відрізняються від істот, — писав колись Демієн Бро́дерік, — інакше вони не варті навіть паперу, на якому їх прописали».[6] І справді — Конструктор життя сам по собі не є істотою: еволюція така ж безтілесна, як і юдео-християнське божество. Він присутній у кожному закуточку Природи — аж до жовтавого листочка, що ось-ось впаде. Його масштаби планетарні. Йому мільярди років. Непроявлений, Він природно постає із законів фізики. Усе перелічене можна сказати про Бога. Правда ж?

6. Сумнівно, що ця цитата належить Демієну, оскільки прямої згадки твору, з якого вона взята, не вдається знайти. — Прим. пер.

І все ж у Творця не знайшлося місця для розуму, Йому і не вдалося сотворити собі тіло. Якось так вийшло, що за людськими мірками плоди Його праці геть кепської якості, внутрішньо суперечливі. І в більшості вони *далекі від красивих*.

Можна сказати, що Дарвін *відкрив* Бога — Бога, що не вписувався у традиційні теологічні очікування, тому його відкриття лишилося непоміченим. Якби Дарвін дізнався, що життя створене розумним агентом — безтілесним розумом, який любить усіх нас (а хто осмілиться сказати інакше, того враз настигне блискавка), — тоді люди б сказали: «Господи, та це ж Бог!».

Однак Дарвін відкрив дивного, чужого Бога — не просто зручного «невимовного», а *дійсно відмінного від нас*. Якби Еволюція була Богом, Ним був би не Єгова. Ним був би Азатóт[7] з усесвіту Говарда Лавкрафта — сліпий недоумкуватий Бог, що під тихі монотонні схлипи флейт хаотично клубочиться у центрі всесвіту.

Якби ви подивилися на Природу *незашореним оком,* зрозуміли б, про що я.

А втім, богомольці ж неабияк вірять у якесь туманне божество. І хто *справді* вірив у невиразне божество, одразу мав би впізнати свого дивного творця, коли Дарвін вигукнув: «Он воно що!».

Ось вам і щира віра у те, що Наука колись знайде Бога. І Наука дійсно знайшла до якоїсь міри божественного творця людського роду, однак про нього віряни і слова не хочуть чути. Адже вони очікували почути про *свого* Бога, вельми конкретного Бога, якого хочуть бачити всюди. Що ж, хай чекають далі, позаяк велике відкриття *уже сталося* — Дарвін знайшов Азатота.

Тим краще для нас, хочу сказати! Мені подобається, що я можу перехитрити нашого Творця. Та й не думаю, що комусь сподобається бути чиїмсь хом'ячком. Я радий, що переможцем виявився Азатот, а не Одін.

---

7. https://uk.wikipedia.org/wiki/%D0%90%D0%B7%D0%B0%D1%82%D0%BE%D1%82

Диво еволюції

Еволюція працює. І вже це — диво!

Ні, я серйозно: якщо вам подобається дивуватися, еволюція справді варта вашого подиву.

Як у цьому всесвіті взагалі могла виникнути оптимізація? Якщо Природу сконструював розумний агент, хто ж тоді створив того агента? Де шукати першу модель без творця? Загадка криється навіть не в тому, як зробити перший етап самозавантаження надефективним і надоптимізованим. Як це *взагалі* стало можливим?

Цю нескінченну регресію і розв'язує еволюція. І не тому́, що вона сконструйована надрозумним агентом чи надефективна. Попри свою недолугість та неефективність, вона все ж працю́є. Хіба *це* не диво?

Професія часто зобов'язує мене говорити про повільність, довільність і сліпоту еволюції. І після моїх промов хтось обов'язково заводить: «*Ви* от кажете, що еволюція не здатна до декількох змін водночас, що через довільність мутацій еволюція незграбна. Хіба це не риторика *креаціоністів*? Вони теж кажуть: якщо потрусити коробку з деталями, годинника ніяк не зібрати».

На це креаціоністам слід відповідати не тим, що таким чином якось *можна* зібрати годинника. Відповідь: еволюція працює *не* так. Якщо ви мислите еволюцію *як* вихор, що конструює літаки, вітаю: креаціоністи підсунули вам солом'яне опудало замість еволюції.

Справжня відповідь: складні механізми виникають або шляхом поступових змін, або через пристосування вже наявних складних систем до нових цілей. Бі́лки можуть перестрибувати з крони на крону лише завдяки силі своїх м'язів, а дальність стрибка вже визначається аеродинамічними можливостями їхнього тіла. Так і з'явилися політухи — їхня аеродинамічна будова дає змогу планерувати на невеликі відстані. Якби всі птахи

вимерли, нащадки політух за десять мільйонів років цілком собі могли б зайняти їхню екологічну нішу, розвинувши літальні перетинки у крилах. А якийсь креаціоніст на це відреагував би якось так: «Яка користь від половинчастих крил? Махнув — і відразу розбився. Птахобілки розвивалися поступово — це ж треба яка нісенітниця!».

Ось так з однієї складної адаптації постає нова, а потім ще одна, позаяк складність може наростати поступово, починаючи з однієї мутації.

Першим у гру вступає простий, але принаймні *трохи* корисний сам по собі ген А. Він швидко поширюється й стає звичним елементом генофонду. До пари йому приходить ген В, корисний тільки у присутності гена А, що вже надійно закріпився у генофонді, тому під тиском добору ген В теж має надійні шанси на закріплення. Згодом виникає видозмінений варіант А*, що стає залежним від гена В, і ця взаємозалежність зберігається між генами А та А*. Вітайте ген В, залежний від генів А*, В і В*, причому останній залежний від генів А* та В. Невдовзі ви отримаєте «незвідно складний» механізм: варто вилучити з нього хоча б один елемент — і вся система руйнується.

Уявіть собі цей довжелезний потяг аж до найдрібнішого елемента: *не* ламаючи всього механізму, ви можете послаблювати залежність однієї деталі від іншої. Повторивши процедуру кілька разів, врешті зможете вийняти цілу деталь, не пошкодивши механізму, аж поки не розберете складний механічний годинник до його примітивної сонячної версії.

Наприклад, ДНК зберігає інформацію у стабільному форматі, що дає змогу точно відтворювати її. Рибосома перетворює цю збережену інформацію у послідовність амінокислот (білóк), котрі складаються у різноманітні хімічно активні форми. Так, об'єднана система ДНК і рибосом здатна збудувати будь-які білкові механізми. Але яка користь від тієї ДНК, якщо немає рибосом, які перетворюють її інформацію на білки? Який сенс у рибосомах без ДНК, що задає їм інформацію для синтезу потрібних білків?

Не завжди організми лишають по собі скам'янілості. Не завжди еволюціоністи можуть відстежити покроковий шлях розвитку організмів. Але в цьому випадку ми *дійсно* знаємо, як це відбулося. РНК має спільну з ДНК властивість поширювати інформацію і самовідтворюватися, хоч РНК менш стійка і продукує менш точні копії. Крім того, РНК може згортати білки у хімічно активні форми, хоч цей процес і менш стабільний, ніж формування амінокислотних ланцюгів білків. Майже напевно, РНК — єдиний ген А, що передує взаємозалежним А* та В.

Важливо не лише те, що РНК *кепсько* справляється з функцією, яку ДНК виконує разом із білками, а й те, що вона взагалі здатна цю функцію виконувати. Дивовижно, що *однісінька молекула* здатна і зберігати інформацію, й управляти хімічними реакціями. А там уже не так і важливо, що вона не може виконувати ту роботу сама.

Яким був найперший ген-реплікатор? Можливо, то була нитка РНК, адже *за дивним збігом обставин* її хімічні складники могли утворитися на Землі ще до появи життя близько чотирьох мільярдів років тому. Не забуваймо: еволюція *не* пояснює, звідки взялося життя; еволюційна біологія *не здатна* пояснити перший реплікатор, бо на те він і *перший*, що йому не передує якийсь інший. Еволюція описує статистичні тенденції у процесах відтворення. А перший реплікатор — справа чистого випадку, а не статистики. Уявлення, що еволюція мусить пояснити *початки* життя, — солом'яне опудало, більше схоже на перекручування креаціоністів.

Якби ви мали нагоду спостерігати за доісторичним хімічним супом у момент появи першого реплікатора, який назавжди змінив Землю, навряд чи вас вразила б *точність* його відтворення. Мавпа під наркотою, може, відтворила б його навіть краще. Найперше відтворення не мало жодних ознак ретельності, властивої сучасним реплікаторам, адже перший утворився *випадково*. Не було *потреби* точно відтворювати нитку РНК, хімічний гіперцикл чи візерунок на глині. Таке просто мало статися. І навіть якщо розглядати цю подію як ізольовану, більш

імовірною за надзвичайно малоймовірну вона не стане. Така подія могла статися лише *раз*, хоч і припливних басейнів було чимало. А через кілька мільярдів років ці реплікатори уже ступили на поверхню Місяця.

Перший випадковий реплікатор став найважливішою молекулою в усій історії життя. Але якщо ви заходитеся наділяти її всілякими чудотворними можливостями, що сприяють відтворенню, ви *вказуєте у хибному напрямку*.

Не треба думати, що у політичній боротьбі між креаціоністами й еволюціоністами шанувальники еволюції обов'язково на стороні науки. У науки дуже точне уявлення про спроможності еволюції. Віддаючи еволюції більше, ніж вона може, ви не «воюєте на стороні еволюції» проти креаціонізму. Ви стоїте на хибних, з погляду науки, засадах. Крапка. Наполягаючи на тому, що вихор *може* зібрати літак, ви ставите ногу на міну, закладену креаціоністами. Еволюція така розумна! Славімо її! Глядіть: я прихиляю коліно на вірність науці! Що улесливіше я говорю про еволюцію, то завзятіше воюю на її боці.

Наділяти еволюцію тим, чого вона не робить, — знецінювати *справжнє* диво: не у *майстерності* конструювання, а в самому природному *виникненні* процесів конструювання.

Зрештою, пропоную відмовитися від ідеї, що еволюція — неперевершений конструктор чи поводир видів, якого люди мали б наслідувати. Взоруватися нам на еволюцію як на ідеал конструювання — що вимагати від сучасної складної бактерії повернутися до біохімічних шляхів найпершого доісторичного реплікатора. Недарма Томас Генрі Хáкслі, на прізвисько Дарвінів Бульдог, казав:

> «Зрозуміймо раз і назавжди: етичний прогрес суспільства полягає не у взоруванні на космічні процеси (і поготів не у втечі від них), а в боротьбі з ними».[1]

1. Thomas Henry Huxley, *Evolution and Ethics and Other Essays*.

Хакслі — чи не останній, кому можна дорікнути у спротиві еволюції. Він добре розумів її, тому й написав те, що написав.

В еволюцій мозок від курки, а витівки — на віки

У попередньому есеї я писав:

> «У науки дуже точне уявлення про спроможності еволюції. Віддаючи еволюції більше, ніж вона може, ви не "воюєте на стороні еволюції" проти креаціонізму. Ви стоїте на хибних з погляду науки засадах. Крапка».

Тут пишу про знані сла́бощі та рамки еволюцій. Так, саме «еволюцій», позаяк еволюція лисиці працює всупереч еволюції кролика, жодна з них не перетинається з еволюцією змій, тому ні в кролика, ні в лисиці немає отруйних іклів.

Хай там я пишу про недоліки еволюції, від того не зробився засланим агентом креаціоністів. Візьміть базовий підручник з еволюційної біології (583[1]— на рівень складніший, якщо вам треба вивести рівняння). Еволюції, навіть зі своїми пояснювальними обмеженнями, і далі роз'яснюють біологічні явища; ба більше — саме ці межі й потрібні, щоб їх збагнути. Не забули, що еволюції дивовижні не ефективністю, а простою дієвістю?

Людський інтелект настільки складний, що ніхто не знайшов путнього способу обчислити його ефективність. Хоч природний добір і не простий для пізнання, він куди *простіший за* мозок людини, а ще — повільніший, менш ефективний, як і личить першому в історії оптимізаційному процесу. Насправді еволюційні процеси достатньо прості, аби ми могли навіть підраховувати, *наскільки вони примітивні*.

Еволюції ще й повільні. Наскільки? Уявімо сприятливу для реплікації мутацію, що підвищує пристосованість на 3%, — тобто в середньому носії цього гена залишають у 1,03 раза більше потомства, ніж ті, хто його не має. Припустимо і те, що

1. https://web.archive.org/web/20071122200230/http://www.tiem.utk.edu/~gavrila/583/583.html

мутація передається. Скільки мине часу, поки вона розповсюдиться на всю популяцію? Залежить від розміру популяції. Щоб ген, що підвищує пристосованість на 3%, передався популяції зі ста тисяч особин і закріпився у генофонді, у середньому знадобиться 768 поколінь. Для популяції у п'ятсот тисяч — 875 поколінь. Загальна формула має вигляд:

$$\text{Кількість поколінь для закріплення} = 2\ln(\mathbf{N}) / \mathbf{s},$$

де **N** — розмір популяції, а (1 + **s**) — пристосованість. (Якщо кожен носій цього гена має у 1,03 раза більше нащадків, ніж неносій, то **s** = 0,03.)

Тому, якщо чисельність популяції становить близько мільйона особин — приблизно стільки її було за часів суспільства мисливців і збирачів, — то для закріплення в генофонді гена, що дає одновідсоткову перевагу, знадобиться близько 2 763 поколінь.[2]

Немає чого дивуватися. Гени поширюються самі по собі. Ніякої феї Еволюції, яка спостерігає за генофондом: «Хммм, отой ген передається доволі швидко — треба передати його всім». У світі ринкової економіки, якщо чиїсь інвестиції окупаються на 20% — особливо коли за цим стоїть прозорий механізм, — його капітал швидко росте від ще більших інвестицій, і завжди знайдуться охочі повторити цей успіх. Гени ж передаються геть інакше: уявіть, що Генрі Форду довелося зібрати один автомобіль, продати його, купити запчастини лише для 1,01 нової машини, продати їх — і повторювати цю схему знову й знову, аж поки не назбирається мільйон автівок.

У першу чергу гени посилюють свою присутність у генофонді. Ось це рівняння простіше і не залежить від розміру популяції:

$$\text{Імовірність закріплення} = 2\mathbf{s}$$

2. Graur, D. and Li, W.H. 2000. *Fundamentals of Molecular Evolution*, 2nd edition. Sinauer Associates, Sunderland, MA.

Мутація, що дає 3% переваги (а за еволюційними мірками це надзвичайно багато), має лише шестивідсотковий шанс поширитися — принаймні в цьому випадку.[3] Такі мутації можуть виникати неодноразово, але в популяції з мільйоном особин і точністю копіювання близько 10^{-8} помилок на основу за покоління на ще одну нагоду доведеться чекати сотні поколінь — і навіть тоді ймовірність її закріплення лишатиметься на рівні тих самих 6%.

І все ж у **довгостроковій перспективі** еволюція таки досягає свого. (Про це повсякчас і писатиму.)

Для *складних* адаптацій потрібно *дуже-дуже* багато часу. Першою з'являється алель A — корисна сама по собі; у генофонді вона закріплюється тисячу поколінь. *Тільки* потім приходить алель B, що залежить від A, і стає на шлях закріплення. Якщо середовище існування організму *статистично схиляється* до нижчих температур, хутро стає для нього помітною перевагою. Гени ж частково формують середовище для інших генів: якщо ген B залежить від гена A, B не даватиме суттєвої переваги, якщо A ще не *закріпився* у тому середовищі.

Скажімо, ген B забезпечує 5% переваги лише за наявності гена A, а без нього не дає жодної. Якщо ж ген A трапляється лише в 1% популяції, то ген B приносить користь лише в одному зі ста випадків — отже, його середня пристосувальна перевага становить 0,05%, а ймовірність закріплення — близько 0,1%. А коли йдеться про складнішу адаптацію, *перший* ген A закріплюється приблизно за тисячу поколінь, *тоді* ген B — ще за тисячу, *тоді* ген A* — ще за тисячу... і ось уже за кілька мільйонів років з'являється нова складна адаптація.

А інші еволюції не наслідують тієї адаптації: якщо зміїна еволюція розробляє новий вид отрути, ця особливість ніяк не допомагає еволюції лисиць чи левів.

А програміст *за день* може створити новий механізм із сотнею взаємозалежних компонентів. Як таке взагалі *можливо*? Не

3. Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection. *IV. Proc. Camb. Philos. Soc.* 23: 607-615.

певен, чи зможу відповісти на це питання, і припускаю, що наука так само не зможе, адже людський розум куди складніший за еволюції. Я міг би порозводити руками: «Знаєте, цілеспрямований зворотний логічний ланцюжок з опертям на комбінаторні модульні репрезентації», — але так чи інакше вам не вдасться створити свою людину. Разом з тим люди вміють заздалегідь розробляти нові деталі, тримаючи в умі подальші, вносити скоординовані одночасні зміни у взаємозалежні механізми, навчатися зі спостережень за окремими тестовими випадками, зосереджуватися на проблемних місцях, міркуючи над їх розв'язком, а також вирішувати, які коригування варто спробувати першими, а не чекати на божу милість. За мірками природного добору усе вище — ніщо інше, як чиста магія.

Люди вміють робити таке, на що еволюції не спромоглися б *за весь очікуваний період* існування всесвіту. «Один випускник за годину зробить те, що еволюції не подужають і за мільярди років», — довелося мені якось почути за обідом від біологині Си́нтії Ке́ньйон. Сучасні знання з біології дають нам зрозуміти, що еволюції винаходили рухоме колесо загалом *тричі*.

Не забувайте, що програміст ще й зберігає частину коду в інтернеті.

Безумовно, деякі результати праці еволюції гідні подиву навіть у порівнянні з її найкращою технологією — *Homo sapiens*. Але наш кембрійський вибух тільки розпочався. Ми збираємо знання не так давно... Стривайте... Останні чотириста років? А де в чому біологія й досі перевершує найкращі людські технології: ми ще не навчилися створювати самовідтворювальну систему розміром із метелика. В іншому наші технології розбивають біологію вщент: колесо, сталь, рушниця, ніж, кіл, ракета, транзистор, атомна електростанція. З кожним десятиліттям шальки терезів дедалі більше схиляються на бік людства.

Повторюся: покладатися на природний добір як на зразок інженерного мислення — те саме, що вимагати від сучасної складної бактерії відтворити біохімічні шляхи першого незграбного реплікатора. Запустити його у сучасне висококонкурентне

довкілля — його б одразу зжерли. Та сама доля чекала б і на проєктувальника, який тестує випадкові точкові мутації і чекає 768 ітерацій на результат, що принесе покращення на 3%.

Тому й кажу: не віддавайте еволюціям більше, ніж вони заслуговують.

У наступному есеї розкажу ще більше цікавих зв'язків між математикою та еволюцією!

Еволюції не стосуються ані корпорацій, ані нанопристроїв

> «Ні фізичні закони, ні математичні правила ніколи не переставали застосовуватися до реальності. Тому я вірю, що еволюція триває й досі. І тому я вірю, що природа — хтось схарактеризував її як кров на кігтях і пазурях — переходить на наступний етап...»
>
> «[Виключити дарвінівську еволюцію] — що намагатися усунути гравітацію. Доки ресурси обмежені, а індивіди конкурують за передачу своїх характеристик, доти існує селекційний тиск».
>
> — Перрі Мецґер (прогнозує, що природний добір пануватиме до невизначеного майбутнього)

Як і багато інших галузей, еволюційну біологію варто мислити кількісно, а не якісно. Психічна сила була б сприятливою мутацією, що «поширюється іноді, але не завжди»? Чи слід очікувати, що вона поширюється? Ось це приклад мислення якостями: якщо X істинне, то істинне і Y; якщо психічні сили корисні, вони можуть передаватися. В есеї «В еволюцій мозок від курки, а витівки — на віки» я розповідав про рівняння ймовірності закріплення сприятливої мутації (та, яка дасть на 3% більше потомства, передасться з імовірністю 6%). Тільки завдяки такому кількісному підходу ми розуміємо: мутації, що *рідко дають користь*, поширюються з украй малою ймовірністю, а складні адаптації практично не мають шансів сформуватися, якщо ними не користуються постійно. Якби люди справді мали психічні сили, їх би використовували всі й постійно. Не лише тому, що вони були б надзвичайно корисними, а й тому, що в іншому разі вони б у нас не еволюціонували.

«Доки ресурси обмежені, а індивіди конкурують за передачу своїх характеристик, доти існує селекційний тиск». Це — приклад мислення якостями. *Скільки* селекційного тиску?

На звання найважливішого рівняння в еволюційній біології претендують декілька кандидатів, але я візьму рівняння Прайса[1], яке у найпростішому формулюванні має вигляд:

$$\Delta z = \mathrm{cov}(v_i, z_i)$$

(Середня характеристика в популяції змінюється відповідно до коваріації між значенням характеристики та відносною пристосованістю індивідів.)

Дуже потужна загальна формула. Візьмімо певний ген зросту — назвімо його Z (характеристика, що змінюється). Згідно з рівнянням Прайса, зміна ймовірності передачі цього гена визначається коваріацією гена з репродуктивною пристосованістю. Або ж розглянемо *зріст загалом* як характеристику Z, незалежно від окремих генів. У такому разі рівняння Прайса стверджує, що зміна середнього зросту в наступному поколінні дорівнюватиме коваріації між зростом і відносною репродуктивною пристосованістю.

(Принаймні доти, доки зріст передається напряму. Коли поліпшується харчування, фіксований генотип дає вищий зріст — отже, до рівняння Прайса варто додати поправковий член. Якщо між багатьма генами простежуються складні нелінійні взаємодії, варто або додати поправковий член, або вирахувати рівняння якимось таким складним способом, що воно перестане бути зрозумілим.)

Багато що прояснюється, коли розглянути різні варіанти й виведення рівняння Прайса. Наприклад, остаточне рівняння розкриває те, що середня характеристика змінюється відповідно до її *коваріації* з *відносною* пристосованістю, а не *абсолютною*. Тобто якщо ген Фродо рятує весь свій вид від загибелі, середнє значення цієї характеристики не зросте, адже його вчинок дав

1. https://en.wikipedia.org/wiki/Price_equation

однакову користь усім генотипам і не *коваріював* із *відносною* пристосованістю.

Кажуть, Джордж Прайс настільки переймався тим, що його рівняння має погані наслідки для альтруїзму, що скоїв самогубство. А може, у нього були інші проблеми. (У цьому блозі ніхто не закликає вас накладати на себе руки після ознайомлення з Прайсовим рівнянням.)

Міркуючи над рівнянням, ви можете прийти до думки, що «обмежені ресурси» та «індивіди, що конкурують за передачу своїх генів» — *недостатні* умови для еволюції. Не достатньо й «істот, що самовідтворюються», і навіть «конкуренції між організмами, що відтворюються».

Чи еволюціонують корпорації? Безумовно, вони конкурують. Час від часу навіть породжують дочірні підприємства. Ресурси у них обмежені. Іноді вони ще й помирають.

Але наскільки дочірні підприємства схожі на материнські? Значна частина внутрішньої культури корпорації йде від начальства, хоч гендиректори і *не* розмножуються шляхом поділу. Прайсове рівняння працює доти, доки характеристики передаються між поколіннями. Якщо праправнуки вже мало схожі на своїх прапрабатьків, то це всього лише результат чотирьох поколінь добору — а риси ще давніших предків тим паче зникають із геному. Безперечно, внутрішня культура корпорації може впливати на її дочірні структури, але це зовсім не те саме, що передача ДНК — яка кодується дискретно, а не безперервно, і передається з похибкою близько 10^{-8} на нуклеотид за покоління.

ДНК передає спадкову інформацію на *мільйони* поколінь — настільки складні адаптації можуть виникнути внаслідок лише еволюції, адже тривалість дискретної ДНК достатня для того, щоб ген, що передає 3% переваги, поширився на 768 поколінь. Якби корпорації і реплікували себе з цифровою точністю, у світі РНК вони б протрималися не довше десяти поколінь.

Певна річ, корпорації проходять *добір* у тому сенсі, що некомпетентні корпорації банкрутіють. З цього логічно виходить, що має бути більше рис корпорацій, які вкладаються у

компетенцію. За аналогією, будь-яка зоря, яка стає новою незабаром після її формування, з меншою ймовірністю буде помітною на нічному небі. Та якщо за примхою динаміки зоряного горіння одна зоря живе довше за іншу, це не означає, що ймовірність тривалішого горіння її нащадків зростає — така риса не передається іншим зорям. Ми не станемо очікувати, що астрофізики у майбутньому відкриють *складні* структурні характеристики зір, завдяки яким вони горітимуть довше. Для такої механічної адаптації потрібен набагато більший *накопичувальний* тиск, ніж одноразовий відсів.

Згадаймо принцип, який я описав ув есеї «Зухвалість Айнштайна»: величезний обсяг даних, необхідний для створення загальної теорії відносності, привів Айнштайна до *одного конкретного рівняння*. У порівнянні з цим кількість даних, потрібна, щоб із навмисно розглянутої можливості дійти впевненості у 99,9%, суттєво менша. За аналогією, складні риси корпорацій, для яких потрібно вказати сотні бітів, існують завдяки людському розуму, а не низці поколінь, що проходять через низькоточну еволюцію. Біологія показує: мутації геть випадкові, а еволюція постачає тисячі бітів сумарного селекційного тиску. У корпораціях же люди розробляють тисячобітні складні «мутації», а подальший селекційний тиск у вигляді питання «Чи корпорація збанкрутіла?» додає ще кілька бітів до пояснення ваших спостережень.

Очікується, що передові молекулярні нанотехнології — штучний різновид, а не біологія — зможуть самовідтворюватися протягом тисяч поколінь із цифровою точністю. Чи закріпиться тут Прайсове рівняння?

Кореляція — це коваріація двох випадкових величин, поділена на добуток їхніх стандартних відхилень: якщо за характеристикою A можна з високою ймовірністю передбачити характеристику B, вони можуть сильно «корелювати», навіть якщо A коливається від 0 до 9, а B — від 50,0001 до 50,0009. Прайсове рівняння побудоване на *коваріації* характеристик з розмноженням — не кореляції! Стиснувши дисперсію характеристик до

тоненької стрічки, ви побачите, як коваріація іде на значний спад. Те саме відбувається і з накопичувальними змінами у характеристиках.

Серед іншого, в аналітичному центрі «Форсайт Інстіт'ют» (Foresight Institute[2]) пропонують зашифровувати інструкції реплікації будь-якого нанопристрою так, щоб навіть перекидання одного біта в цих інструкціях спричиняло *повне* перемішування всього розшифрованого виходу. Якщо виготовляти всі нанопристрої як точні молекулярні копії — причому помилки на конвеєрі, якими б вони не були, не передаються спадково, адже нащадок щоразу отримує цифрову копію первинних зашифрованих інструкцій, щоб виготовляти власних нащадків, — тоді для еволюції у таких нанопристроїв просто не залишиться простору.

За що справді варто перейматися, то це пріони — самовідтворювальні виробничі хиби, що не випливають із зашифрованих інструкцій. Скажімо, роботична рука не захоплює атом карбону, потрібний для складання її гомолога, — і внаслідок цього рука нащадка теж не здатна захопити атом карбону, навіть якщо самі зашифровані інструкції залишилися абсолютно незмінними. Але наскільки ймовірна *кореляція* між таким способом передавання помилки та *вищим* коефіцієнтом відтворення? Нехай один нанопристрій виробляє одну копію себе кожні тисячу секунд, а новий нанопристрій, на диво, ефективніший (у нього не просто пріон, а *сприятливий* пріон) — він самовідтворюється уже за 999,99999 секунди. І для цього, як бачите, потрібно лише відокремити один атом карбону. Оскільки під час відтворення дисперсія невелика, відповідно коваріація теж залишається малою.

Тож як часто мають відтворюватися ті нанопристрої? Якщо тільки у їхньому розпорядженні не знайдеться більше атомів, аніж у всій Сонячній системі (або, якщо вже на те пішло, у видимому всесвіті), то перш ніж вони стикнуться з нестачею ресурсів, мине лише кілька поколінь. «Обмежені ресурси» самі по собі не

2. https://foresight.org/about/

є достатньою умовою для еволюції; щоб вивільнити їх, у значній частині популяції має регулярно хтось помирати. Насправді «покоління» — це не стільки ціле число, скільки інтеграл від частки популяції, що складається переважно з новостворених індивідів.

Ось це і лякає мене найдужче у сірому слизі або нанотехнологічній зброї: поглинувши всю Землю, на планеті не залишиться нічого, крім *власне слизу*, і жодне інше цікаве життя зародитися не зможе. Оскільки діамант стабільніший за білок, що тримається купи вандерваальсівським зв'язком, слизу знадобиться перезібрати деякі зі своїх складників лише тоді, коли на Землю впаде астероїд. Навіть якби про пріони можна було говорити як про потужну підтримку еволюції — а вона й так повільна і з цифровою ДНК! — між остаточним поглинанням Землі слизом і смертю Сонця пройшло б, мабуть, менше одного покоління.

Підсумуємо: якщо у вас наявні *всі* наступні властивості:

- частинки, що відтворюються,
- значна дисперсія у їхніх характеристиках,
- значна дисперсія у їхньому процесі відтворення,
- стійка кореляція між характеристиками і відтворенням,
- високоточна далекосяжна спадковість характеристик,
- часте відтворення значної частини репродуктивної популяції,
- *усі* ці компоненти лишаються істинними впродовж *багатьох* ітерацій,

тоді з'являється значний *накопичувальний* селекційний тиск, достатній, щоб еволюція могла породжувати складні адаптації.

Доеволюціонувати до вимирання

Теза про те, що еволюція працює на благо її видів, — *доволі* розповсюджена омана. Пам'ятаєте розповіді про двох кроликів, які дають потомство у вісім особин, чим «докладаються до виживання виду»? Сучасний еволюціоніст і не подумав би сказати щось подібне — радше він сам схрестився б із тим кроликом.

Уже не вперше вам доводиться тримати в голові кілька абстрактних понять одночасно і чітко розмежовувати їх. Еволюція не *діє* на окрему особину, адже кожен організм живе з тим генетичним набором, з яким народився. Еволюція розгортається на рівні відтворюваних популяцій, і для прояву її наслідків потрібен час. Людям часто здається, що якщо якась фея Еволюція *діє на рівні* видів, то вона нібито має *оптимізовувати* на користь виду. Насправді ж еволюція зводиться до змін у частотах генів — частота зростає або зменшується не через користь для всього виду. Наприкінці есею побачимо, що вид цілком може собі доеволюціонувати до вимирання.

Чому дівчат і хлопців народжується приблизно однакова кількість? (Не беремо до уваги божевільні країни, де використовують технології штучного відбору статі.) Чому це спостереження має дивувати? Нехай один чоловік може запліднити двох, десятьох чи сотню жінок. З цього не випливає так, що для виживання виду потрібна рівна кількість чоловіків і жінок. Це спостереження видається ще дивнішим для більшої частини царства тварин, де самці мінімально докладаються до вирощування потомства. Люди ж, з погляду батьківського вкладу, вирізняються навіть серед приматів. Збалансоване співвідношення статей присутнє навіть у видів, де самці запліднюють самок і зникають назавжди.

Уявімо дві групи по різні сторони гори. У таборі А кожна мати народжує двох хлопчиків і двох дівчаток, у таборі Б — трьох дівчаток й одного хлопчика. В обох групах однакова

кількість дітей, однак у таборі Б буде на 50% більше онуків і на 125% більше правнуків. Певно, вам спадає на думку, що у групи Б є значна еволюційна перевага.

Але зверніть увагу на інше: що рідше народжуються хлопчики, то вищою стає їхня репродуктивна цінність — не для *групи* загалом, а для *конкретних* батьків. У кожної дитини двоє батьків обох ста́тей. Тоді у кожному поколінні загальний генетичний внесок чоловіків дорівнює загальному генетичному внеску жінок. Що менше чоловіків, то більший особистий генетичний внесок кожного чоловіка. Якщо усі жінки роблять усе заради блага групи і народжують одного хлопчика на десять дівчат, ви можете зірвати генетичний *куш* — народити лише хлопчиків, і в кожного з них в середньому буде вдесятеро більше онуків, ніж у їхніх двоюрідних сестер.

Отже, хоча груповий добір може підштовхувати до переважання дівчат, індивідуальний добір прагне рівномірного репродуктивного внеску в обидві статі. Якщо подивитися на статистику народжуваності з пологового будинку, з першого погляду видно, що у *Homo sapiens* кількісний баланс між груповим та індивідуальним доборами сильно перехиляється на користь другого.

(Технічно — не зовсім з першого. В індивідуальному доборі надається перевага рівномірному *батьківському вкладу* у нащадків обох статей. Якщо народження та/або виховання хлопчиків удвічі дешевше, то для еволюційно стабільної рівноваги народжуватиметься удвічі більше хлопчиків, ніж дівчаток. Якби в популяції загалом народжувалося порівну хлопчиків і дівчаток, а поява на світ хлопчика обходилася б удвічі дешевше, ви знову зірвали б генетичний куш, народжуючи більше хлопчиків. Отже, статистика з пологових будинків має відображати баланс батьківських витрат на виховання дитини тієї чи іншої статі у суспільстві мисливців-збирачів, причому ви мали б якось це оцінювати. А знаєте: не схоже, щоб для первісної сім'ї виховання дівчинки було *надто* затратним з погляду втрачених

репродуктивних можливостей. Тому здається дивним, що хлопчиків народжується стільки ж, скільки й дівчат.)

Природний добір не стосується ні груп, ні видів, ні навіть *особин*. У видах, що розмножуються статевим шляхом, організм не еволюціонує — він живе з тими генами, з якими народився. Особина — це одноразовий набір генів, що ніколи не повториться. Як із цього взагалі можливий якийсь добір? Коли ви усвідомлюєте, що усі ваші пращури зрештою померли, стає ясно, що теза «виживають найпристосованіші» геть не коректна. А от теза «відтворюються найпристосованіші» точніша, хоч і технічне поняття пристосованості *визначається* суто в межах відтворення.

Частота генів — ось що дійсно стосується природного добору. Щоб отримати складну адаптацію — механізм із багатьма залежними частинами, — кожен новий ген у процесі еволюції узалежнюється від інших генів, що вже надійно закріпилися у генетичному середовищі. А для цього їм потрібна висока частота. І що складніша та машина, то вищою має бути частота. Справжня ознака природного добору — це збільшення частоти певного гена від 0,00001% до 99% генофонду. Це — власне інформація в сенсі теорії інформації. І саме така зміна частот генів необхідна для появи масштабних, складних еволюційних адаптацій.

У процесі добору справжня боротьба не в суперництві організмів за ресурси, адже всі учасники зникнуть уже в наступному поколінні. Справжня складність у суперництві але́лів за частоту у генофонді. Ось він — той довготривалий наслідок, що створює стійку інформацію. Двоє баранів, що бекають і б'ються рогами, — лише проміжні тіні у цьому процесі.

Якийсь алель цілком може поширитися аж до закріплення, перевершивши альтернативний алель, що працював «заради блага виду». Якби Летюче Локшинне Чудовисько за допомогою магії створило вид із гендерним складом, ідеально оптимізованим для виживання *виду* — відновлення після подій, що загрожують вимиранням, освоєння нових ніш тощо, — еволюція

швидко зруйнувала б цей видовий оптимум до індивідуального: самці та самки зрештою робили б рівноцінний батьківський внесок.

Уявімо «ген Фродо», що жертвує своїм носієм *для збереження цілого виду* від вимирання. Що це означає для частоти його алелів? Вона іде на спад. На цьому все.

Якщо загроза видового вимирання періодично повторюється (назвімо такі умови «середовищем Ба́ффі»), ген Фродо систематично ітиме на спад і невдовзі зникне, а з ним — увесь вид. Приклад умоглядний? Можливо. Якщо людство не планує змінювати свою біологічну форму на найближче століття, є сенс у тому, щоб почати клонувати Ганді.

У середовищі вірусів присутній тиск між швидким розмноженням окремих вірусів та вигодою, що дає живий організм хазяїна для передачі хвороби. Ось це — реалістичний приклад групового добору. Якщо вірус еволюціонує у плані пристосованості аж до того, що тиск групового добору не перевищує тиск індивідуального добору, вірус незабаром зникне. Не знаю, чи якусь хворобу колись виявляли на стадії еволюції до вимирання, але, ймовірно, таке траплялося кілька разів.

Гени, що порушують сегрегацію, підривають механізми, які зазвичай гарантують рівномірність статевого розмноження. Наприклад, у статевій хромосомі деяких мишей-самців присутній ген, що спотворює сегрегацію, через що народжуються лише самці, причому усі вони стають носіями цього гена. Ці самці запліднюють самиць, які народжують таких же самців, і так цей процес іде далі. «Це ж шахрайство!» — ви можете так сказати, але це лише з погляду людини, адже репродуктивна пристосованість цього алеля настільки висока, що у наступних поколіннях вона відтворює вдвічі більше копій себе, ніж альтернативний без мутацій. Навіть коли чисельність самок падає, самці-носії не стають менш схильними до спарювання, ніж будь-який інший самець, тому алель, що спотворює сегрегацію, залишається вдвічі більш пристосованим, ніж його альтернатива. Припускають, що реалістичний груповий добір міг певною мірою підтримувати

цей ген на такому низькому рівні частоти, на якому його спостерігають. У такому випадку, якби миші розвинули здібність літати і мігрувати у теплі краї на зиму, певно, вони б утворили єдину репродуктивну популяцію й еволюціонували б аж до вимирання, щойно ген-порушник закріпився у генофонді.

Близько 50% геному кукурудзи складають транспозони[1]— ділянки ДНК, чия основна функція зводиться до самокопіювання і перенесення *в інші ділянки*. Виглядає, що клас транспозонів, відомий як P-елементи, уперше з'явився у дрозофіли лише в середині XX століття і поширився по всіх популяціях виду протягом приблизно 50 років. У людському геномі так звана Alu-послідовність, чиї елементи мають довжину триста пар основ, повторюється від трьохсот тисяч до мільйона разів. Можливо, вона не призведе до вимирання виду, але точно не допомагає йому, адже транспозони спричиняють приріст мутацій, які переважно шкідливі і знижують точність ефективного копіювання ДНК. І все ж такі гени-шахраї надзвичайно пристосовані.

Припустимо, що у якогось статевого виду виробився *бездоганно точний* механізм копіювання ДНК. Позаяк більшість мутацій шкідлива, цей генний комплекс становить перевагу для його носіїв. Тепер ви, мабуть, замислилися: якщо корисні мутації такі нечасті, хіба немутабельні організми не опиняться у невигідному становищі? Проте у статевих видах сприятлива мутація, що почалася в мутабельних організмах, може поширитися і на нащадків немутабельних. Тоді як мутабельні організми у кожному поколінні страждають від дегенеративних мутацій, немутабельні можуть отримувати мутації статевим шляхом і вигоду від будь-яких сприятливих мутацій, що виникають у мутабельних організмів. Так мутабельні організми повністю втрачають перевагу. Бездоганний механізм копіювання ДНК збільшує свою частоту аж до закріплення. Десять тисяч років потому світ сковує льодовиковий період, і вид зникає з лиця землі — він доеволюціонував до повного вимирання.

1. https://uk.wikipedia.org/wiki/Транспозон

Ефект свідка[2] проявляється так: коли хтось опиняється в біді, з більшою ймовірністю втрутяться окремі особини, а не вся група. Студенту коледжу з явними ознаками епілептичного нападу надав допомогу окремий свідок у 85% випадків і лише у 31% випадків — 5 свідків. Припускаю, що навіть якщо родинні зв'язки у племені мисливців-збирачів були достатньо міцними, щоб створити добір для виникнення допомоги непрямим родичам (причому якщо у них було кілька потенційних помічників), генетичні перегони озброєнь могли призвести до того, що допомога надходила *лише від останнього*. Кожен зволікає й очікує, що хтось інший зробить цей крок. Са́ме зараз людство спостерігає за численними загрозами вимирання на рівні його ж виду, і мушу сказати: не так багато людей готові щось із цим зробити. Якщо ми програємо цю битву через те, що ніхто просто не вийшов на поле бою, тоді ми просто доеволюціонуємося до вимирання, як і ті численні види, яких ми вже не бачимо.

Ракові клітини почуваються нівроку у тілі хазяїна: процвітають і накопичують все більше ресурсів, значно перевершуючи у чисельності своїх більш слухняних співмешканців. До пори, до часу.

Багатоклітинні організми виживають лише завдяки наявності потужних внутрішніх механізмів, які *протидіють еволюційним процесам* всередині себе. Коли ці клітини починають еволюціонувати, вони швидко доеволюціонують до вимирання — організм просто загине.

Не варто вихваляти еволюцію за турботу, яку вона виявляє особині, — майже усі ваші предки померли. Не варто вихваляти її і за турботу, яку вона виявляє видам, — ніхто ще не знаходив складної адаптації, яку можна було б інтерпретувати лише як таку, що діє задля збереження виду, причому математика еволюції, видно, свідчить, що таку адаптацію практично неможливо знайти. Як бачите, вид цілком може доеволюціонувати до вимирання. І, можливо, саме зараз людство завершує цей процес. Не

2. https://uk.wikipedia.org/wiki/Ефект_свідка_(психологія)

варто навіть вихваляти еволюцію за турботу, яку вона виявляє генам, адже боротьба між двома альтернативними алелями на одній ділянці — це гра з нульовою сумою для частоти.

Пристосованість не завжди на вашому боці.

Трагедія групового добору

До 1966 року цілком звичним ділом було зустріти фахових біологів, котрі відстоювали еволюційні гіпотези, які сьогодні здалися б магічним мисленням. Ті плутані припущення зіграли неабияку роль в історії розвитку теорії еволюції — роль помилки, що призвела до виправлення. Подібним чином дурість англійських королів призвела до появи Великої хартії вольностей і дала початок конституційній демократії.

Як романтичний приклад годиться згадати Ве́ро Ві́нна-Едвардса, Во́рдера Алле́ та Джозефа Ллойда Бре́ретона, які, серед інших, вважали, що хижаки зусиллям волі здатні утримуватися від розмноження, щоб запобігти перенаселенню власного ареалу й подальшому виснаженню популяції здобичі.

Але еволюція не відчиняє воріт перед довільними цілями. Ви не можете пояснювати тріскачку на хвості гримучника користю для тварин, яких за інших обставин він міг би вкусити. Жодна фея Еволюція не вирішує, коли слід поширити той чи інший ген: він сам *має* стати *прямою причиною* власного закріплення у наступному поколінні. Звісно, спостерігаючи за різким спадом популяції лисиць, які з'їли всіх кроликів ув ареалі, ми зі своїм відчуттям гармонії голосимо: «Щось варто було зробити із цим!». Але як генний комплекс, що *обмежує відтворення* — усіх істот! — може спричинити зростання власної частоти в наступному поколінні?

Людина, яка створює невеличку іграшкову екосистему — заради забави, як із моделлю залізниці, — може розчаруватися, якщо штучно вирощені популяції лисиць і кроликів, на які пішло стільки зусиль, знищать самі себе: лисиці поїдять усіх кроликів, а тоді самі загинуть від голоду. Та людина возитиметься зі своєю іграшковою екосистемою (обмежити розмноження лисиць — очевидне рішення, що відразу спадає на думку) доти, доки та екосистема виглядатиме акуратною. Звісно, у Природи немає такої людини. Але коли *це* нас зупиняло? *Ми* ж виходимо

з припущення про *гармонію*, тож нам лишається лише підшукати правдоподібний аргумент, який змусив би Природу хотіти *того самого*, але *еволюційним* шляхом.

Очевидно, добір на рівні особини не призведе до того, що окрема особина обмежуватиме власне розмноження. Особини, що не обмежують себе у розмноженні, очікувано приведуть у світ більше нащадків, аніж ті, що обмежують.

(**Додаток:** індивідуальний добір не спричинить *обмежень у розмноженні окремих особин*. Безумовно, індивідуальний добір може породити особин, котрі, отримавши всі необхідні ресурси, використовують їх для знесення чотирьох великих яєць замість восьми маленьких — *не* для збереження соціальних ресурсів, а тому, що це *оптимальне* співвідношення кількості яєць до ймовірності їх виживання. Однак це не вирішує трагедію спільних ресурсів.)

Уявімо, що популяція виду розпалася на субпопуляції, переважно ізольовані, які зрідка схрещувалися між собою. Згодом субпопуляції, що обмежували власне розмноження, вимирали з меншою ймовірністю, надсилали більше посланців і засновували нові колонії, заповнюючи території вимерлих популяцій.

Такий сценарій не те щоб математично *неможливий* — він *дуже складний*.

Вигоду від обмеженого розмноження отримують не тільки ті, хто обмежує себе. У цьому і полягає засадничість проблеми. Якщо деякі лисиці утримуються від народження дитинчат, яким дістаються впольовані кролики, ті не з'їдені кролики не дістаються *лише* дитинчатам, батьки яких мають адаптацію обмежувального розмноження. Лисиці без таких обмежень (та їхні численні нащадки) будуть раді забрати собі невпольованих кроликів. В умовах такого тиску ген стримування може вижити тільки тоді, коли переваги стримування залишаються переважно за тим, кого він стримує.

Точніше, вимога має вигляд: **C/B < FST**, де **C** — ціна альтруїзму для донора, **B** — вигода альтруїзму для отримувача, **FST** — просторова структура популяції (тобто середня *спорідненість*

між випадково вибраним організмом та його випадково вибраним сусідом, де «сусід» — будь-яка інша лисиця, яка отримує вигоду від стримування альтруїстичної лисиці). (Гадаю, це[1] рівняння похідне від іншого, лише з іншими символами; це найкраще, що я знайшов ув інтернеті.)

Отже, щодо емпіричної просторової структури популяцій лисиць і кроликів — чи справді ціна репродуктивного обмеження достатньо мала, а емпірична вигода від зменшення голоду достатньо велика, щоб аргумент групового добору спрацював?

Якщо вірити математиці, ймовірність цього доволі низька. Наприклад, у цій симуляції[2] ціна для альтруїстів становить 3% пристосованості; у суто альтруїстичних групах вона вдвічі вища, ніж у чисто егоїстичних. При розмірах субпопуляції у 25 особин і 20% заміни померлих посланцями з інших груп наслідок — поліморфізм альтруїзму та егоїзму. Якщо розміри субпопуляції подвояться до 50 особин, закріплюється егоїзм. Якщо для альтруїстів ціна зросла до 6%, закріплюється егоїзм. Якщо вигода для альтруїстів падає вдвічі, закріплюється егоїзм (принаймні для значної частини популяції). Коли ціна альтруїзму перевищує 10%, сусідні групи мусять бути дуже малими — близько п'яти особин, — щоб груповий добір почав працювати. Це якось не переконує, що лисиці обмежуватимуть своє розмноження.

Думаю, тепер ви здогадалися, що прихильники групового добору остаточно програли цей науковий спір. На перешкоді стала не математика, а емпіричне спостереження: лисиці *не* обмежували себе репродуктивно (не пригадаю точну назву виду; ішлося не про кроликів і лисиць), і системи хижак-жертва ламаються весь час. Пізніше груповий добір відродиться у кардинально іншій формі: якщо говорити мовою математики, *існує* така структура сусідства, що створює ненульовий тиск групового добору, *необов'язково* здатний подолати контртиск індивідуального добору, і якщо ви не візьмете цього до уваги, ваші розрахунки будуть хибувати. Крапка. А ще — розвинені механізми

1. https://www.pnas.org/cgi/reprint/72/1/143.pdf

2. https://www.gnxp.com/MT2/archives/003540.html

примусу (від початку не допускалися) повністю змінюють гру. З чого б це вже історична наукова суперечка варта згадки у цьому блозі?

Десятиліття після дискусії одного біолога навідала блискуча ідея. Математичні умови, необхідні для того, щоб груповий добір перевершив індивідуальний, були надто екстремальними, тож їх навіть не виявили в Природі. Чому б тоді не створити ті умови штучно, у лабораторії? Ідею розвинув[3] Майкл Вéйд: він повторно відбирав популяції комах, у субпопуляції яких було мінімум дорослих особин. Що в нього вийшло? Чи ті комахи обмежували себе у розмноженні? Чи жили вони своє спокійне життя у ресурсному достатку?

Ні. Дорослі особини пристосувалися їсти власні яйця й личинки (особливо личинки жіночих особин).

Звісно, добір у невеликих субпопуляціях не призведе до добору особин, які обмежують *себе* у розмноженні. Виживають ті, хто їсть дітей *інших* особин (особливо жіночої статі).

І щойно ви отримуєте результат цього експерименту — у ретроспективі більш ніж очевидний, — раптово стає ясно, що перші прихильники групового добору дозволили романтизму і відчуттю гармонії затуманити свої передбачення природних явищ.

Перед вами архетипний приклад упущеної Третьої альтернативи, що вилилася у раціоналізацію заздалегідь визначеного нижнього рядка, вона ж призвела до хибного виправдання, і вони навмисне припинили пошуки. Прихильники групового добору не завдали собі клопоту розпочати з чистим розумом; прийшовши до ідеї групового добору, вони *нейтрально* екстраполювали ймовірний результат на подальші випадки. Вони починали з прекрасної ідеї: популяція лисиць може добровільно обмежувати себе у розмноженні до тієї міри, щоб кролики встигали відтворюватися, що Природа вміє досягати досконалого балансу. Тому вони шукали підставу, чому так мало статися, і

3. https://www.pnas.org/cgi/reprint/73/12/4604.pdf

прийшли до гіпотези групового добору. Однак, коли зрозуміли, якого результату від групового добору вони *прагнуть*, так і не спромоглися розглянути *менш* ідеальні та менш гармонійні адаптації, які міг би просувати груповий добір. Якби вони *справді* намагалися спокійно і нейтрально передбачити результат добору для малих субпопуляцій, то подумали б про поїдання дітей інших організмів або якийсь подібний «потворний» результат — *задовго до того*, як вони придумали щось настільки еволюційно неординарне, як *індивідуальне обмеження у розмноженні!*

Цей випадок також ілюструє думку, яку я висловлював ув есеї «Зухвалість Айнштайна»: у широчезному просторі можливих відповідей майже всі реальні зусилля зводяться до виокремлення однієї можливої відповіді. Якщо гіпотеза привертає вашу увагу неналежним чином — ваше відчуття прекрасного вказує, якою має бути Природа, тоді як природний добір не передбачає феї Еволюції, яка була б у захваті від вашої ідеї, — сама ця обставина наперед визначить ваш провал, хіба що ви вмієте очищати розум і починати з абсолютно чистого аркуша.

У теорії найдурніший у світі пеньок може говорити про світло сонця, але від того воно світити ніяк не перестане. Навіть якщо відповідь пропонує пришелепа під наркотою, ви повинні нейтрально розрахувати свідчення за і проти, а вже потім, за необхідності, попрощатися з переконанням.

На практиці ж прихильники групового добору були приречені, бо їхній нижній рядок від початку містив відчуття прекрасного, а нижній рядок Природи був написаний природним добором. Обидва результати не мали жодної принципової підстави корелювати, та й у реальності нічого такого не відбувалося. І жодна гнівна суперечка опісля не змінила того.

Якщо ви починаєте з власних бажань, що має робити Природа, враховуєте спостережувані підстави для Її дій, а потім раціоналізуєте надзвичайно переконливий аргумент, чому Природа має прийти до того результату, якого ви від неї хочете (причому на підставах Природи), вона *просто* не слухатиме. Всесвіт не має розуму — йому однаково до хитромудрих політичних

дебатів. Ви можете хоч цілий день переконувати гравітацію направляти воду *вгору*, але вода як текла вниз, так і буде, хай там що ви їй розказуєте. Всесвіт просто глухий. «Природа — це такий найвідданіший фанатик, який вперто і віддано слідує своїм упередженням і навідріз відмовляється поступитися перед найпереконливішими людськими раціоналізаціями», — якось сказав Дж. Р. Моллой.

Я часто раджу друзям вивчати еволюційну біологію, оскільки у сучасній галузі студентів вже вчать уникати раціоналізації та виправляти помилки. Фізиків та інженерів не навчають уникати антропоморфізації електронів, оскільки ті не проявляють жодних ознак розумної поведінки. Природний добір створює видимість задуму, яку через її чужість людям складно осягнути, тому під час навчання студентів-біологів попереджають про це. Це гарне тренування для будь-кого, хто працює головою, але *особливо* важливо для тих, хто прагне ясно мислити про інші дивакуваті процеси, схожі на розумові, що, однак, діють за геть іншими принципами.

Критерій помилкової оптимізації

У попередніх есеях я докладно розглядав форми раціоналізації, завдяки яким наші переконання видаються значно краще узгодженими з даними, ніж це є насправді. І тут я не перебільшую. Якби нам удалося подолати цей засадничий метабіас і безпосередньо побачити передбачення, які випливають з кожної гіпотези, ми змогли б оговтатися від майже будь-якої хиби, пов'язаної з даними.

Дзеркальний виклик для теорії рішень полягає в ідентифікації того, який варіант дій *на ділі* відповідає обраному критерію вибору. Якщо ваші моральні принципи зобов'язують забезпечувати людей ноутбуками, то яка дія *відповідає* їм: купівля інкрустованого коштовностями ноутбука за мільйон доларів чи передача п'яти тисяч ноутбуків відповідній благодійній організації?

Складається враження, що ми набули своєрідного хисту проголошувати, ніби практично будь-яку дію можна виправдати практично будь-якою метою. Теоретик флогістону, що пояснює збільшення маси магнію під час горіння, принципово нічим не відрізняється від інквізитора, який твердить, що безмежна Божа любов до всіх Його дітей допускає спалення декількох із них на вогнищі.

І це далеко не секрет. Політика була рисою середовища предків. Ми — нащадки тих, хто найпереконливіше обґрунтовував необхідність страти ненависного племінного суперника Уґлака. (Тож ми точно не його діти.)

І все ж ... чи можливо *довести*, що якби Роберт Мугабе піклувався *виключно* про добробут Зімбабве, то пішов би з поста президента? Ви можете *стверджувати*, що політична стратегія випливає з мети. Але хіба ми щойно не побачили, що люди можуть припасувати будь-яку мету до будь-якої стратегії? Звідки ви взяли, що ви праві, а Мугабе — ні? (Можна назвати чимало причин, чому це гарна здогадка, але прошу простежити, куди я веду.)

У мотивах людей багато граней і мало ясності; процеси прийняття рішень за складністю не поступаються самому мозку. Крім того, світ сам по собі дуже заплутаний, як і кожен вибір реальної стратегії. А чи можемо ми взагалі *довести*, що люди раціоналізують — систематично спотворюють зв'язок між принципами і стратегією, — якщо у нас немає навіть єдиної точки опори, на яку можна було б твердо спертися? Та й способу дізнатися, до чого *точно* приведе один оптимізаційний критерій, теж немає. (Насправді можна спостерігати, як люди *не погоджуються* з політикою організації у спосіб, що дивним чином корелює з їхніми інтересами, хоча вони й заперечують наявність таких інтересів. Але йдемо далі.)

Де ж той універсальний відкритий консеквенціалістський оптимізатор, в який можна завантажити всю мораль в XML-файл і дізнатися, які рекомендації й моральні принципи *справді* можна застосувати в нашому світі? Чи є бодай один такий реальний випадок, коли ми *точно* знаємо, що саме схвалює критерій вибору? Де ж той *чистий* механізм моральних міркувань — функція корисності, вільна від усіх випадкових бажань, які могли б спотворити її оптимізацію, — результати якої можна зіставити з раціоналізаціями людини з подібною функцією корисності?

Звісно ж, це наш старий добрий друг — чужий бог! Природний добір гарантовано вільний від милосердя, любові, співчуття, уявлень про гармонію, політичних фаворитів, ідеологічних уподобань, академічних амбіцій, лібертаріанських або соціалістичних симпатій, а також суперечок між блакитними та зеленими. Природний добір не *максимізує* свій критерій інклюзивної пристосованості — не настільки він розумний. Та коли ви розглядаєте результати природного добору, неминуче помітите, що вони оптимізовані *лише* для інклюзивної пристосованості, а не в інтересах сільського господарства США.

У випадку еволюційної науки, наприклад, ув есеї «Трагедія групового добору», ми можемо безпосередньо порівняти людські раціоналізації з результатом *чистої* оптимізації за відомим

критерієм. Яким, на думку Вінна-Едвардса, був результат групового добору в масштабі малих субпопуляцій? Добровільне самообмеження у розмноженні й достатня кількість харчів для всіх. Яким був лабораторний результат? Канібалізм.

Тепер ви можете запитати: чи справді ці історії з еволюційної науки стосуються людської моралі, якій начхати на інклюзивну пристосованість, коли вона стає на шляху любові, співчуття, естетики, зцілення, свободи, справедливості тощо? У людей і поняття «інклюзивної пристосованості» не було аж до XX ст.

Але я запитаю у відповідь: якщо ми не можемо чітко побачити результат одного монотонного оптимізаційного критерію — якщо не можемо навіть натренувати себе чути одну чисту ноту, — то як ми тоді збираємося почути оркестр? Як нам зрозуміти, що «Завжди дій егоїстично» чи «Завжди слухай уряд» — кепські принципи, до яких не варто пристосовуватися, якщо навіть *гени, оптимізовані для інклюзивної пристосованості*, породжують організми, що жертвують власною репродуктивною здатністю заради збереження соціальних ресурсів?

Щоб натренуватися бачити ясно, потрібні прості вправи.

Виконавці адаптацій — не максимізатори пристосованості

> «Про окремі організми найдоцільніше мислити як про виконавців адаптацій, а не як про максимізаторів пристосованості».
>
> — «Психологічні основи культури», Джон Ту́бі та Ле́да Косми́діс

П'ятдесят тисяч років тому смакові рецептори *Homo sapiens* підштовхували їх до пошуків найбільш дефіцитних харчових ресурсів — цукру та жиру. Калорій, словом. На сьогодні функційний контекст смакових рецепторів змінився, а самі рецептори — ні. Калорії, яких тепер більш ніж достатньо (принаймні у країнах першого світу), почали завдавати шкоди. Мікронутрієнти, яких ніколи не бракувало у листі й горіхах, відсутні у хлібі, але наші смакові рецептори не жаліються. Ріжок морозива — надстимул, у якому цукру, жиру та солі більше, ніж у всьому, що траплялося предкам у природному середовищі.

Жодна людина з *усвідомленою* метою максимізувати інклюзивну пристосованість алелів ніколи б не з'їла печиво, хіба що помирала з голоду. Але про окремі організми найдоцільніше мислити як про виконавців адаптацій, а не як про максимізаторів пристосованості.

Тостер, хоч розробник навмисне виготовив його для запікання тостів, не містить у собі інтелект розробника — якщо ви спробуєте впихнути у нього цілу буханку, він не переналаштується автоматично під ваш запит. Хрестоподібна викрутка не переформатується у викрутку з плоскою голівкою. Хоча ці знаряддя створили ми, вони існують незалежно від нас.

На атомах викрутки не вигравійовані крихітні XML-теги, що описували б їхнє «об'єктивне» призначення. Безумовно, розробник мав певне бачення того, як має працювати викрутка, але

в реальному світі вона далеко не завжди *реалізується* саме так. Якщо ви забули, що розробник — окрема від виробу сутність, ви можете думати: «*Призначення* викрутки — крутити гвинти» — ніби воно випливає з безпосередньої властивості самої викрутки, а не стану розуму розробника. Можливо, для вас стане несподіванкою, що викрутка не переналаштовується під гвинт із плоскою голівкою, позаяк її *призначення* — крутити гвинти.

Причина появи викрутки — розум розробника, який уявив мислимий гвинт і мислимий пристрій для крутіння. *Власне* функціонування викрутки, її *прилягання* до голівки гвинта, *не можуть* бути об'єктивною причиною існування викрутки — як і майбутнє не може бути причиною минулого. Але мозок розробника — сутність, яка існувала у минулому, — може бути причиною появи викрутки.

А *наслідок* існування викрутки може не збігатися з уявними наслідками у розумі розробника. Лезо інструменту може зіслизнути й порізати руку того, хто ним користується.

Смисл викрутки полягає в тому, що вона існує у розумі того, хто її використовує, а не в крихітних ярличках на її атомах. Можливо, розробник і задумував, щоб вона крутила гвинти. Проте вбивця може використати її як зброю. А потім він її упустить, ту викрутку знайде дитина і, граючись, довбатиме нею дірки у стіні.

Отже, *причина* виникнення викрутки, її *форма*, *наслідок* її появи та різноманітні *смисли* — усе це різні сутності, і тільки *одну* з них можна знайти у самій викрутці.

Звідки беруть початок смакові рецептори? Не від розумного розробника, котрий вимислив наслідки тих рецепторів, а від замороженої історії спадковості: Адаму подобався цукор, він їв яблука і передав свої гени; Барбара любила цукор, їла яблука і передала свої гени; Чарлі любив цукор, їв яблука і передав свої гени — і після 2 763 поколінь цей алель закріпився в популяції. Часом, для спрощення, ми зводимо цю гігантську історію до формулювання «Це зробила еволюція». Але від того вона не

стає миттєвою локальною подією, як-от винайдення викрутки. Ось така вона — *об'єктивна причина* появи смакових рецепторів.

Яка *об'єктивна форма* смакових рецепторів? Технічно — молекулярний сенсор, підключений до схеми підсилення. Оскільки смакові рецептори впливають на пошук їжі опосередковано, вони додають ще один рівень заплутаності. Вони впливають на розум організму, спонукаючи його *хотіти* їжу, схожу на щойно з'їдену.

Який *об'єктивний наслідок* появи смакових рецепторів? У сучасної людини з розвинених країн простежуються декілька ланцюгів причиново-наслідкових зв'язків: бажання їсти більше шоколаду, намір їсти більше шоколаду, поїдання шоколаду, набирання ваги, зменшення кількості побачень, нижчий репродуктивний успіх. Цей наслідок прямо *протилежний* до основної закономірності у довгому ланцюгу успіхів предків, що стала причиною появи форми смакових рецепторів. Та, оскільки переїдання — недавня проблема, ніяка значуща еволюція (зведена закономірність у середовищі предків) ще не вплинула на форму смакових рецепторів.

У чому *смисл* поїдання шоколаду? А він лежить між вами і вашою моральною філософією. Як на мене, шоколад просто смачний, і хотілося б, щоб він був менш шкідливим. Прийнятним рішенням було б переробити або склад шоколаду, або мою біохімію.

Склавши всі міркування докупи, виходить приблизно так: «Сьогодні люди роблять те, що сприяло б поширенню генів у первісному суспільстві, незалежно від того, чи допомагає це їм сьогодні». Але не зовсім так, оскільки ми *насправді* не питаємося, яка саме поведінка максимізувала інклюзивну пристосованість наших предків. Сьогодні ми робимо те, чого наші предки й уявити не могли. У первісному суспільстві ще не придумали шоколад.

Тому варто розглядати наші смакові рецептори як *адаптацію* до умов життя предків, серед яких були голодування,

яблука, смажені кролики. А сучасні люди *користаються* тими ж адаптаціями у нових умовах, де є дешевий шоколад і постійний вплив реклами.

Тому й кажуть: про окремі організми найдоцільніше мислити як про виконавців адаптацій, а не як про максимізаторів пристосованості.

Еволюційна психологія

Як «IRC-чат» і «TCP/IP-протокол», словосполучення «репродуктивний орган» тавтологічне. *Усі* органи репродуктивні. Як у птахів з'явилися крила? Фея Пташина еволюція вирішила, що вміння літати стане птахам у великій пригоді? У птахів розвинулися крила, бо вони посприяли розмноженню їхніх предків. Не стояли осторонь і серце, і легені, і геніталії. Щонайбільше можна розрізняти *безпосередньо* репродуктивні та *опосередковано* репродуктивні органи.

Це спостереження справедливе і для мозку — найскладнішої біологічної системи органів. Деякі структури головного мозку беруть безпосередню участь у репродуктивній функції (наприклад, ті, що пов'язані з хіттю), тоді як інші залучені до неї опосередковано (наприклад, через регуляцію гніву).

Звідки у людей взявся гнів? Фея Людська еволюція вирішила, що він може стати людям у пригоді? Нейронна структура гніву — це репродуктивний орган, як і ваша печінка. У виду *Homo sapiens* гнів зберігся, тому що гнівливі предки мали більше дітей. *Іншого способу його появи просто не існує.*

Це *історичне* спостереження про походження гніву збиває з пантелику багатьох людей. «Стривайте. То, як на вас, коли я гніваюся, підсвідомо хочу дітей? Це точно не те, що приходить на думку, коли мені розбивають носа», — кажуть вони.

Ні. *Ні. **Ні. НІ!***

Про окремі організми найдоцільніше мислити як про виконавців адаптацій, а не як про максимізаторів пристосованості. *Причина* адаптації, її *форма* і *наслідок* — усе це різні сутності. Створюючи тостер, ви не очікуєте, що він перебудується лише тому, що вам раптом забаглося засунути в нього цілу буханку. Так, ви мали намір підсмажити тост, але ваш намір стосується вашого ментального стану, а не тостера. У тостера немає відчуття власного призначення.

Тостер не належить до класу об'єктів, яким властиво мати наміри. Оскільки він узагалі не є розумом, немає підстав приписувати йому цілеспрямованість. Якщо *ми* бачимо у тостері цілеспрямованість, то не вважаємо, що він знає про це, бо не вважаємо, що тостер знає *хоча б щось*.

Ця ситуація нагадує класичний тест на називання кольору шрифту, яким написано слово «блакитний». Учасникам потрібно більше часу, щоб назвати колір, оскільки їм треба відплутати значення літер від їхнього кольору. Приміром, зі словом «вітер» таких проблем не виникне.

Проте людський мозок — поряд із тим, що він є артефактом історичного еволюційного процесу, — водночас виступає розумом, здатним породжувати власні наміри, смисли, бажання, цілі й плани. Що бджола, що людина є продуктами еволюції, але розробником може бути лише людина. За аналогією, бджола — це «вітер», а людина — «блакитний».

Когнітивні та еволюційні причини *онтологічно різняться*: когнітивні ґрунтуються на нейронах, еволюційні — на предках.

Найбільш явна форма когнітивної причини — свідома, що проявляється в намірах і планах, наприклад: піти в магазин чи засмажити тост. А от емоція має ще й фізичний прояв у вигляді потоку нейронних імпульсів або поширюваної хмари гормонів. Те саме стосується інстинкту, польоту уяви чи придушеної думки: просканувавши мозок, побудувавши його тривимірну модель та розшифрувавши код, ви могли б їх *виявити*.

Навіть підсвідомі процеси мають фізичний прояв у мозку. «Влада розбещує», — спостеріг якось Джон Да́льберг-А́ктон. Хтозна, чи Сталін бачив себе альтруїстом, чи прагнув найбільшого блага для більшості (може, і ні). Але складається ймовірне враження, що десь в його мозку, у нейроструктурах, функціонував механізм підсилення задоволення від влади та детектор очікування посилення чи послаблення влади. Якби в мозку Сталіна нічого не корелювало з владою — жодного мерехтіння при відчутті політичної сили та відсутності сигналу від політичної

слабкості — як тоді ми могли б дізнатися, що влада розбестила його мозок?

Тиск еволюційного добору *онтологічно відмінний* від біологічних артефактів, що їх створює. Еволюційною причиною появи крил у птахів є те, що мільйони пташиних предків із поступово покращеними крилами відтворювалися частіше (зі статистичною закономірністю), ніж їхні конкуренти, у результаті чого ця риса закріпилася. Ми стискаємо цю велетенську історико-статистичну макроподію до твердження «Це зробила еволюція».

Природний добір онтологічно відмінний від істот; еволюція — це не маленька хутряна істота, що ховається у незвіданому лісі. Еволюція — це причинова статистична закономірність у репродуктивній історії предків.

Ця логіка прикладається і до мозку. Еволюція створила ляскі крила, але не розуміє ляскоту, створила ноги, але не розуміє ходьби. Вона створила кістки з іонів кальцію, але кістки не мають чіткого уявлення про міцність (а про інклюзивну пристосованість і поготів). Еволюція створила мозок, який здатен творити, але в того мозку розуміння еволюції було не більше, ніж у птахів — розуміння аеродинаміки. Аж до XX ст. у жодному людському мозку не простежувалося ні йоти складного абстрактного розуміння *інклюзивної пристосованості*.

Тож коли кажуть «Еволюційна мета гніву — підвищити інклюзивну пристосованість», спостерігається тенденція зісковзувати до «*Мета* гніву — відтворення», а там до «Когнітивна мета гніву — відтворення». Ні! Статистична закономірність історії предків не живе *в* мозку, так само як наміри не живуть у тостері!

Думати, що ваша вбудована структура гніву недвозначно втілює потяг до відтворення, — те саме, що думати про руку як про втілене бажання хапати предмети.

Не те щоб ваша фізична рука повністю відокремлена від ментальних бажань. У конкретних випадках ви можете рухати пальцями зусиллям волі. Зігнувшись, щоб підняти пенні, ви

здійснюєте акт волі, але він відрізняється від того акту, завдяки якому виросла ваша рука.

Слід розрізняти одноразову подію конкретного гніву (гнів-1, гнів-2, гнів-3) від засадничої нейронної структури, що породжує гнів. Подія гніву — це когнітивна причина, і сама подія може бути когнітивно зумовленою, хоч ви й не хотіли, щоб структура гніву була підключена до мозку.

Отже, слід розрізняти подію гніву, нейронну структуру гніву, генний комплекс, що заклав основу для цього нейронного шаблону, та історичну макроподію, що пояснює присутність цього генного комплексу.

Якщо й існує дисципліна, що дійсно *вимагає* надприскіпливого погляду, то це еволюційна психологія.

Любі читачі, до вашої уваги непристойна і кумедна історія. Якось у барі перестрілися чоловік із жінкою. Чоловік помітив її ніжну шкіру і пружні груди, що у предків могли повідомляти про фертильність. Тут же на нього вплинули макіяж і бюстгальтер. Але це аніскілечки його не хвилює — зрештою, йому просто подобається жінка. Його нейронна структура, що ідентифікує ніжну шкіру, поняття не має про свою мету — виявлення ознак фертильності (так само і на атомах його руки не закарбовані XML-теги у вигляді «<ціль>хапати предмети</ціль>»). Жінці запали його впевнена посмішка і рішучий тембр, що сигналізують про високий статус, а в середовищі предків означали здатність здобувати ресурси для потомства. Вона планує скористатися протизаплідними, хоч її детектори впевнених посмішок знають про це не більше, ніж тостер знає про намір підсмажити тост. Жінку геть не обходять філософські сенси цього бунту, адже її мозок прихистив у себе креаціонізм, і вона палко заперечує еволюцію. Його так само не переймають ніякі філософські сенси цього бунту, бо він просто хоче добре провести ніч. Вони приїздять у готель, роздягаються. Чоловік надягає презерватив. Дітей він не хоче — чисто бажання покохатися, спричинене дофаміном і норадреналіном. П'ятдесят тисяч років тому, як не було ніяких презервативів, таке бажання майже завжди

призводило до зачаття. Вони кохаються, приймають душ і йдуть собі — хто куди. Основний об'єктивний наслідок — прибуток для бару, готелю та виробників презервативів, хоч ні чоловік, ні жінка не мали наміру збагачувати перелічені бізнеси. І це не має жодного стосунку до статистичних закономірностей відтворення предків п'ятдесят тисяч років тому, які пояснюють набуття ними відповідних генів і поведінкових схем.

Для коректного розуміння еволюційної психології слід тримати в голові багато заплутаних, але тісно пов'язаних абстрактних тверджень, при цьому важливо чітко їх розрізняти і не змішувати.

Особливо витончений еволюційно-психологічний експеримент

> «В одному канадському дослідженні 1989 року дорослих респондентів просили уявити смерть дітей різного віку та оцінити, у яких випадках батьківське відчуття втрати було б найсильнішим. Результати, подані на графіку, демонструють зростання інтенсивності горя до підліткового віку дитини, після чого цей показник починає спадати. Коли цю криву зіставили з іншою, що відображала зміни у репродуктивному потенціалі упродовж життєвого циклу (закономірність, вирахувана з демографічних даних Канади), між ними виявили напрочуд високу кореляцію. Утім, ще сильнішою — майже ідеальною — була кореляція між кривою переживання горя у цих сучасних канадців і кривою репродуктивного потенціалу у людей мисливсько-збиральницького суспільства — племені кунґ в Африці. Іншими словами, вікова динаміка горя майже повністю відповідала дарвіністським прогнозам, побудованим з урахуванням демографічних реалій середовища еволюційних предків людини ... Кореляція у першому випадку становила 0,64, тоді як у другому вона була значно вищою — 0,92».
>
> (Роберт Райт, стислий виклад «Людське горе: чи пов'язана його інтенсивність з репродуктивною цінністю померлих?»[1])

Застереження: цю статтю я не читав, бо а) не знайшов її у мережі, б) вона не стосується моєї нинішньої роботи

1. "Human Grief: Is Its Intensity Related to the Reproductive Value of the Deceased?" Crawford, C. B., Salter, B. E., and Lang, K.L. *Ethology and Sociobiology* 10:297-307.

безпосередньо. Втім, судячи з процитованого опису, виглядає як серйозно продуманий експеримент. [Ось тут[2] версія з обмеженим доступом. Дякую Бе́нджі Фалленшта́йн. Дивно, що я її не знайшов. Оце читаю зараз ... перевіряю основні розрахунки: кореляції ті самі, розмір вибірки N=221.]

Найочевидніша *грубість* цього дослідження, як бачите, полягає у його методології: респондентів просили уявити батьківську скорботу замість того, щоб опитувати реальних батьків, які мають дітей відповідного віку. (Імовірно, за таких умов абсолютні значення оцінок були б вищими, а частка охочих узяти участь у дослідженні — нижчою.) Водночас, наскільки мені відомо, отримані результати загалом узгоджуються з даними більш безпосередніх досліджень батьківського горя, у яких аналізували інші типи залежностей, зокрема «сиру́» кореляцію між інтенсивністю батьківського горя та віком дитини.

Разом з тим подивіться на добре пропрацьовані сторони експерименту:

- Кореляція — 0,92! Така підозріло висока — як на еволюцію — кореляція може навести на питання: хіба міг добір настільки точно налаштувати цей процес? А потім приходить усвідомлення, що тиск добору не просто *тонко відрегулював* батьківське горе, а, по суті, *витіснив його з самого початку*.
- Ті, хто стверджують, що еволюційна психологія не висунула жодних прогнозів, як би іронічно це не звучало, показують себе жертвами синдрому «ніхто не знає, чого не знає наука». Якби не еволюційна психологія, ви б і не *подумали про можливість проведення такого експерименту*.
- Цей експеримент — один із найвитонченіших прикладів, які мені доводилося бачити, — наочно ілюструє різницю між наявністю *свідомого або підсвідомого прихованого мотиву* та *реалізацією адаптації, що функціонує без будь-якої чутливості до вихідного тиску добору, що колись зумовив її формування*.

2. https://www.sciencedirect.com/science/article/abs/pii/016230958990006X

Батьківське горе — *навіть на підсвідомому рівні* — не спрямоване на оцінку репродуктивної цінності — інакше воно впливало б на її оновлення для канадських батьків, а не для батьків з племені кунґ. Горе — це адаптація, яка в сучасних умовах просто функціонує: воно присутнє в психіці й продовжує відтворюватися за інерцією.

Батьки переймаються своїми дітьми *не* через репродуктивний внесок. Батькам важливі діти як такі; *некогнітивна еволюційно-історична* ж причина того, чому в людській психіці сформувалася така цінність, полягає в тому, що саме діти є носіями батьківських генів.

Безперечно, саме еволюція є причиною появи будь-яких розумів у нашому всесвіті. Тож, сподіваюся, тепер ви розумієте, чому я хотів провести чітку лінію через мій цинізм щодо прихованих мотивів на цьому еволюційно-когнітивному кордоні. Інакше я так само міг би, стоячи у черзі в супермаркеті, сказати касирові: «Агов, ви всього лише коректно обробляєте візуальну інформацію з метою максимізації власної інклюзивної генетичної пристосованості».

- Здається, 0,92 — найвища кореляція, яку я тільки бачив у еволюційно-психологічних і власне психологічних експериментах. (Втім, мені доводилося бачити й ще вищі значення — до 0,98, — зафіксовані в експерименті, де одну групу респондентів запитували: «Наскільки схожі між собою А і Б?», а іншу — «Яка ймовірність А за умови Б?». Учасникам пропонували запитання на кшталт «Яка ймовірність того, що з бочки, у якій містяться 800 червоних і 200 білих кульок, буде витягнуто 60 червоних і 40 білих кульок?».)

Оскільки тут зібралися самі баєсіанці, ми можемо взяти до уваги наші пріори і спитати, чи, бува, не з чистої випадковості взялися принаймні *деякі* з тих неочікувано високих кореляцій. Можливо, як даність варто прийняти сам еволюційний процес налаштовування, з огляду на надзвичайно потужний тиск

добору, про який ідеться. Решта причин підозріло низької дисперсії — а) чи здатна велика група дорослих респондентів коректно відтворити ступені батьківського горя (вочевидь, ні) і б) чи є плем'я кунґ, яке збереглося до наших днів, *типовим* представником мисливсько-збиральницьких суспільств у цьому контексті. У такому разі постає питання: чи мусила б варіативність *між* типами мисливсько-збиральницьких племен бути настільки великою, щоб дозволити кореляцію 0,92?

Навіть за скептичних апріорних припущень кореляція величиною 0,92 за вибірки N = 221 становить надзвичайно сильне емпіричне свідчення. Відповідно, апостеріорні оцінки щодо цих питань мають бути значно менш скептичними.

- Ви могли подумати, що грубість експерименту полягає в його *потенційному* характері — у вимірюванні уявного, а не ретроспективного, тобто реального, батьківського горя. Проте *са́ме потенційно уявне* горе і *стримує* батьків від поведінки, яка могла б призвести до загибелі дитини. З погляду еволюції, справжня мертва дитина — незворотна втрата, адже еволюція «хоче», щоб батьки навчалися на пережитому болі, а не заподіювали його знову, не поверталися до гедоністичної базової установки і компенсували втрату шляхом виховання інших дітей.
- У подібний спосіб і графік, що корелює з інтенсивністю батьківського горя, відображає не *сукупні незворотні затрати на виховання дитини, яка дожила до певного віку*, а її *очікуваний майбутній репродуктивний потенціал за умови досягнення цього віку*. (Чи була б кореляція *ще вищою*, якби ми враховували лише потенційну репродуктивну ціну виховання дитини віку X до репродуктивної зрілості, повністю ігноруючи всі незворотні затрати, понесені до цього віку?)

Загалом люди схильні фіксувати увагу на незворотних затратах: імовірно, це або адаптація, що обмежує надто часту зміну стратегій (свого роду компенсація імпульсивного полювання за можливостями), або ж афективна реакція — спалах болю, який

ми переживаємо, усвідомлюючи марнування вкладених ресурсів.

Еволюція ж діє інакше. Річ навіть не в тому, що їй «байдуже до незворотних затрат», — вона і близько про них не «думає». «Еволюція» — це просто загальний факт про реальні історичні репродуктивні наслідки.

Тож адаптація батьківського горя сформована так, що вона практично не чутлива до минулих інвестицій у дитину, зате жорстко прив'язана до майбутніх репродуктивних наслідків її затрати. Природний добір не переймається незворотними затратами так, як це робимо ми.

Водночас механізм батьківського горя й далі функціонує так, ніби батьки живуть у середовищі, подібному до племені кунґ, а не в сучасній Канаді. Цю різницю більшість людей змогла б розпізнати.

Що люди, що природний добір не сповні розуму, хоч і кожен — у *власний, стабільний і заплутаний спосіб.*

Надстимули та занепад західної цивілізації

Щонайменше троє людей[12] померли від того, що цілими днями безперервно грали в онлайн-ігри. Вони обміняли своїх партнерів, роботу і дітей на World of Warcraft. Якщо люди мають право грати у відеоігри — а складно уявити право фундаментальніше, — ринок неминуче відповість максимально *захопливими* іграми з усіх, які лише можна продати; у підсумку ж найзатятіші гравці просто вибудуть із генофонду.

Як узагалі може існувати продукт настільки *захопливий*, що навіть після п'ятдесяти семи годин безперервного використання споживач радше взаємодіятиме з ним ще годину, ніж поїсть чи поспить? (Уже чую заперечення: може, той споживач раціонально вирішує віддати перевагу StarCraft ще годинку замість усього подальшого життя. Але прошу: не робіть цього. Будь ласка.)

Солодкий батончик — це *надстимул*: у ньому міститься більше цукру, солі і жиру, ніж у будь-якій їжі, яку тільки могли знайти предки. Батончик ідеально вписується у смакові уподобання, сформовані в тих умовах, і водночас смакує значно краще за будь-яку поживу, яку могли відшукати збирачі, а мисливці — вполювати. Сигнал, який колись надійно корелював з поживною їжею, було зламано й переформатовано до точки в просторі смаків, що ніколи не траплялася у нашому тренувальному наборі даних, — до надвіддаленої аномалії на графіках харчових тенденцій первісних популяцій. Смак, раніше визначений через еволюційні кореляти з користю для здоров'я, було реконструйовано й бездоганно поєднано зі штучно створеними речовинами. На жаль, не існує симетричного ринкового механізму, який змушував би виробників робити смачну їжу водночас і корисною.

1. https://news.bbc.co.uk/2/hi/technology/4137782.stm

2. Перші два покликання, які лишив автор, не доступні. — Прим. пер.

Зрештою, у нас просто немає смакових рецепторів, що безпосередньо реагували б на користь.

Уже відоме відео[3] про еволюцію продукції Dove демонструє ретельно сконструйований приклад іншого надстимулу: на пересічну жінку наносять макіяж, її фотографують професійні фотографи, а згодом додатково обробляють зображення у графічних редакторах, перетворюючи його на образ, гідний рекламного білборда. У такий спосіб створюють стандарт привабливості, з яким у реальному світі — без макіяжу та цифрової обробки — не може зрівнятися жодна жінка. А справжні жінки, конкуруючи з моделями (до прикладу, супермоделі вживають кокаїн, щоб не набирати вагу), яких навіть не зустрінеш на вулиці, вчиняють самогубства.

Так само буденна реальність сильно програє відеогрі — хай навіть гру сприймають через монітор, — так що вона *залучить* гравця настільки, що він не робитиме перерв на їжу чи сон, поки буквально не помре. Хоча мені відомі далеко не всі прийоми, які використовують розробники відеоігор, деякі з них можна впевнено окреслити: завдання з динамічною складністю, у критичних моментах збалансованою між легкою та майже недосяжною; переміжне підкріплення; безперервний позитивний зворотний зв'язок у формі набирання балів або досягнень; а також соціальне залучення, характерне для багатокористувацьких ігор.

Чи існує взагалі межа, на якій ринковий рушій перестає стимулювати створення дедалі захопливіших ігор? Можна сподіватися, що цей механізм зупиняється там, де гравці втрачають роботу: зрештою, без підписки особливо не пограєш. На цій підставі навіть можна вибудувати оптимальне пояснення ігрової залежності: мовляв, задоволення розподіляється за дзвоноподібною кривою, а лише кількох невдахам не пощастило втратити роботу. Станом на 2007 рік поодинокі смертельні випадки, пов'язані з безперервною грою у World of Warcraft протягом

3. https://www.youtube.com/watch?v=iYhCn0jf46U

п'ятдесяти восьми годин, *і справді* виглядали радше винятками, а не правилом. Проте ніхто не скасовував жорсткої конкуренції між розробниками відеоігор. Якщо ви здатні підвищити залученість гравців у свій продукт на 5%, то цілком імовірно, що зможете перетягнути до себе 50% авдиторії конкурентів. Саме так і без того складна проблема *загострюється*.

Якщо люди мають право піддаватися спокусі — а в цьому, власне, й полягає сенс свободи волі, — варто очікувати, що ринок відповідатиме дедалі солодшою спокусою з усіх, яку лише можна продати. Зробити стимули хоча б на 5% привабливішими за пропозиції чинних лідерів — сутність цього ринкового рушія. І цей відсоток неминуче зростатиме, крок за кроком виходячи за горизонт стимульних очікувань наших предків, допоки стимул не перетвориться на надстимул. Поцікавтеся, як змінювалися стандарти продажу жіночої краси з реклами 50-х років XX ст. Як наочно демонструє приклад солодкого батончика, ринковий механізм не зупиняється навіть тоді, коли надстимул починає завдавати споживачеві побічної шкоди.

Тож... Чому б нам просто не сказати ні? Засновкове припущення економіки вільного ринку полягає у тому, що в умовах відсутності насилля і шахрайства люди завжди можуть відмовитися від шкідливої угоди. (Якби це справді було б так, вільний ринок був би не просто *найкращою* з економічних систем, а системою з поодинокими — або й узагалі відсутніми — недоліками.)

Якщо організм систематично пропускає їду, він невдовзі помре — дарма що деякі затяті гравці дізнавалися про це у такий жорстокий спосіб. У середовищі предків навіть зазвичай корисна (а відтак і спокуслива) дія в окремих ситуаціях могла виявитися шкідливою. І люди мають надзвичайно потужну здатність вловлювати ці конкретні випадки за допомогою абстрактного мислення. З іншого боку, ми можемо навигадувати безліч неіснуючих наслідків для окремих випадків, коли духи предків забороняють їсти цілком нормальних кроликів.

Виглядає, що еволюція знайшла компроміс або нашарувала нові системи на старіших підвалинах. Вид *Homo sapiens* і досі легко спокушається їжею, однак завдяки розвиненій префронтальній корі ми водночас маємо *обмежену* здатність чинити опір цій спокусі. Ця здатність не є необмеженою: цілком можливо, що деякі наші предки, наділені надзвичайною силою волі, могли голодувати аж до смерті, приносячи себе в жертву богам, або ж утримуватися від сексуальних контактів надто довго. Так само гравці, далебі, тренували у себе силу волі (у певному сенсі), щоб так довго обходитися без їжі і сну. Ось вам і еволюційна загроза самоконтролю.

Опір спокусі вимагає свідомого розходу вичерпних запасів ментальної енергії. Тому не дуже *чесно* казати, що можна «*просто* відмовитися» — принаймні не без затрат для себе. Навіть люди, які у генетичній лотереї виграли силу волі й завбачливість, все одно платять ціну за опір. Просто їм цю ціну платити легше.

Наша обмежена сила волі еволюціонувала в умовах спокус, із якими стикалися наші предки; у контексті, радикально відмінному від мисливсько-збиральницького середовища, вона може виявлятися значно менш ефективною. Навіть там, де нам вдається протистояти надстимулу, сила волі може вичерпатися значно швидше, ніж в умовах, в яких наші предки опиралися спокусам.

Чи саме́ публічне представлення надстимулів погано відображається навіть на людях, які вміють відмовляти? Чи потрібно тоді забороняти рекламу шоколадних батончиків чи магазинні вивіски з яскравими написами «Морозиво»?

З самої лише наявності проблеми не випливає (без подальших обґрунтувань і значної доказової бази), що держава може розв'язати її. Працівники регуляторних органів зазвичай спрямовують зусилля не на продукти, що поєднують відносно невелику шкоду для споживача зі спокусливими надстимулами, а на ті, чиї недоліки опинилися в центрі публічних скандалів і на газетних шпальтах. Водночас із можливої *неспроможності*

держави щось виправити не випливає, що відповідна проблема *не* має кепських наслідків.

Залишу вам останній аргумент на основі вигаданого свідчення. У своєму онлайн-романі «Після життя» (After Life[4]) Са́ймон Фанк, серед інших сюжетних поворотів, описує заплановане вимирання біологічних *Homo sapiens*, причиною якого стають не армія роботів, а штучні діти — миліші, вродливіші, приємніші у вихованні порівняно з біологічними. Можливо, демографічний крах розвинених суспільств пов'язаний з тим, що ринок пропонує набагато спокусливіші альтернативи на заміну дітовирощування, тоді як заміна підгузків не перетворюється на хоч трохи приємніший досвід. Де ж ті білборди з рекламою «РОЗМНОЖУЙСЯ»? Хто платитиме консультантам з іміджу за те, щоб вони переконували дорослих людей ростити норовливого підлітка замість відпустки на Таїті?

«Врешті-решт, — пише Саймон Фанк, — людський вид було просто перемелено ринковими механізмами».

4. https://interstice.com/%7Esimon/AfterLife/

На уламках бога

Аж до XX століття жодна людина поняття зеленого не мала про «інклюзивну пристосованість» — єдину і непохитну одержимість сліпого недоумкуватого бога. У нас немає вродженої відрази до презервативів чи орального сексу. Перед тим, як нагородити нас сексуальним задоволенням, наш мозок — найголовніший репродуктивний орган — не здійснює жодних перевірок на репродуктивну ефективність.

Чому ж так? Чому ми свідомо *не* одержимі інклюзивною пристосованістю? Чому фея Людська еволюція створила мозок, який винайшов контрацепцію? «*Скільки там того діла*», — думає людина, яка може спроєктувати нову складну систему після обіду.

Як ми вже знаємо, фея Еволюція одержима інклюзивною пристосованістю. Коли вона вирішує, які гени сприяють поширенню, здається, вона не враховує *нічого, крім* кількості копій, що їх виробляє ген. (Воістину дивно!)

Але якщо вже творець інтелекту настільки одержимий, чому б йому не створити розумних агентів — не можемо називати їх людьми, — які дбали б виключно про власну інклюзивну пристосованість? Такі агенти займалися б сексом лише з метою розмноження й не погоджувалися б на секс з контрацепцією. Їхню харчову поведінку визначала б не приємність смаку, а суто раціональне переконання, що їжа необхідна для виживання й репродукції; відповідно, солодкі батончики вони споживали б лише за прямої загрози життю чи відтворенню. Жінки після менопаузи опікувалися б дітьми доти, доки не втратили б працездатність настільки, що подальше лікування вимагало б непропорційних ресурсів, після чого добровільно вкорочували б собі віку.

Виглядає як очевидне покращення дизайну — принаймні з погляду феї.

Тепер стає зрозуміло — як ми й обговорювали вчора, — наскільки складно сконструювати по-справжньому потужного консеквенціаліста. Природний добір певною мірою діє за принципом наслідків, але тільки за умови їх *реального прояву*. Саме тому теоретики людської еволюції змушені вдаватися до надзвичайно витончених абстрактних міркувань, щоб *реконструювати* зв'язки між окремими адаптаціями та репродуктивним успіхом.

Водночас людський мозок *здатен* виображувати ці зв'язки у білкових структурах мозку. То чому ж, створюючи людей, фея Людська еволюція взагалі завдавала собі клопоту з *будь-якою* мотивацією, окрім прямої оптимізації інклюзивної пристосованості?

Відтоді, як білковий мозок уперше сформулював концепцію природного добору, минуло менш ніж два століття. А «інклюзивна пристосованість» — концепт іще витонченіший, іще абстрактніший. Важить не кількість спільних генів. Зрештою, із шимпанзе вас поєднує 95% генів. Має значення спільна *варіативність всередині* відтворюваної популяції: ваша сестра споріднена з вами наполовину, оскільки будь-які варіації у вашому геномі — в межах людського виду — з імовірністю 50% спільні для вашої сестри.

Тільки минулого століття — скоріше навіть за останні п'ятдесят років — еволюційні біологи справді почали розуміти повний спектр причин репродуктивного успіху, як-от: обопільний альтруїзм та затратне сигналювання. Без цього надзвичайно деталізованого знання розумний агент, що ставить собі за мету «максимізувати інклюзивну пристосованість», неминуче піймає облизня.

Чому б тоді не вбудувати це знання у білковий мозок заздалегідь? Чому у нашій голові немає *запрограмованого* поняття «інклюзивної пристосованості» разом із бібліотекою покрокових стратегій? Тоді ми змогли б обійтися без підсилювачів. Організм від народження знав би, що жирна їжа з великою ймовірністю підвищує пристосованість. Якби згодом він засвоїв, що ця

стратегія втратила актуальність, то просто припинив би споживати жирну їжу. Усю систему можна було б перепроєктувати. Тоді не виникло б ані презервативів, ані печива.

Складається враження, що таке, в принципі, цілком можливо. Часом я натрапляю на людей, далеких від розуміння консеквенціалізму, які говорять: «Але якщо в організму немає окремого імпульсу поїсти, він помре з голоду і не зможе передати свої гени». Доки організм знає *це* і має функцію корисності, що максимізує відтворення, він неминуче їстиме автоматично. Власне, за *таким* консеквенціалістським принципом природний добір спроєктував істот, що їдять автоматично.

А як щодо допитливості? Чи консеквенціаліст допитливий лише тоді, коли він бачить якусь особливу підставу для цієї емоції? Чи не призвело б це до втрати важливих знань, що з'являються без явних причин і потребують розвідки? Адже тоді на розвідку він вирушав би лише за наявності знання про її доцільність. Якщо у допитливості ви вбачаєте мотивацію для людини — будучи перебірливою, вона реагує на певні риси проблем, — тоді ця складна адаптація являє собою суто результат консеквенціалістського принципу роботи ДНК — *неявного* відтворення знання. Пращури, котрі проводили подібного роду розвідки, залишали по собі більше нащадків.

Тож, в принципі, суто репродуктивний консеквенціаліст можливий. У принципі, усю історію предків, *опосередковано* відображену у когнітивних адаптаціях, можна перетворити на *безпосереднє* знання, яке вже можна записати на жорсткий диск консеквенціаліста.

Проте сліпий придуркуватий бог не настільки розумний. Еволюція не зводиться до роботи програміста-людини, який одним махом може реорганізувати всю архітектуру коду. Не може він сісти і видруковувати інструкції зі швидкістю шістдесят слів за хвилину.

За мільйони років до гомінідного консеквенціалізму було навчання підсиленням. Винагороджувальні сигнали надійно корелювали з відтворенням. Навряд чи ви просите гомінідний

мозок передбачити, чи дитина, яка їсть жирну їжу, переживе зиму. Тому ДНК і створила білковий мозок, що посилає сигнал винагороди за споживання жирної їжі. А там уже нехай сам організм дізнається, яка здобич смачніша.

ДНК конструює білковий мозок, винагороджувальні сигнали якого у *довгостроковій* перспективі корелюють з репродуктивною пристосованістю, а в *короткостроковій* — з поведінкою організму. Вам не конче розуміти, що цукровмісна їжа восени дає запас калорій на зиму, завдяки якому ви доживете до весни, знайдете пару і до літа зачнете дитину. Яблуко просто припало до смаку — і мозок починає міркувати, як дістати більше яблук із того самого дерева.

Так в організмів розвивається система винагородження за їжу, побудову гнізда, відлякування суперників, допомогу родичам, відкриття значущих істин, формування міцних союзів, переконування суперечками і, звісно ж, секс ...

Коли в гомінідів почали з'являтися мізки, здатні до перехресного консеквенціалістського мислення, вони застосовували його не до абстрактної «інклюзивної пристосованості», а до пошуку вже *наявних* підсилювачів. То була відносно проста хитрість порівняно з цілковитою перебудовою «максимізатора інклюзивної пристосованості». Білкові мізки думали над тим, як роздобути калорії і секс, хоч і не мали найменшого розуміння про «інклюзивну пристосованість».

«Отакої! Я щойно витворив консеквенціаліста! — сказав би інженер. — Тепер я можу зібрати усі свої тяжко здобуті знання про поведінки, що підвищують пристосованість, і чітко прописати їх! Я можу перетворити увесь цей складний механізм навчання підсиленням у просте декларативне знаннєве твердження "Жирна їжа і секс зазвичай підвищують вашу інклюзивну пристосованість". Консеквенціалістське мислення автоматично подбає про все інше. І не призведе до очевидної помилки у вигляді винайдення презерватива!».

Зрештою, інженер і не спроєктував би сітківку в зворотному напрямку.

Сліпий придуркуватий бог не переслідує єдиної мети — його увага розсіяна. Лисиці еволюціонують, бо їдять кроликів; кролики еволюціонують, бо тікають від лисиць. Скільки видів, стільки й еволюцій. Але у *межах* кожного виду сліпий бог одержимий суто інклюзивною пристосованістю. Жодна риса — чи навіть виживання — не цінується доти, доки вона не підвищує репродуктивну пристосованість. Немає жодного сенсу в істоті зі сталевою шкірою, якщо ця риса дасть на 1% менше репродуктивної здатності.

І все ж, коли сліпий придуркуватий бог створив білкові комп'ютери, його мономаніакальна одержимість пристосованістю передалася не зовсім точно. Його оптимізаційний критерій не відтворився як слід. Ми, вислід еволюції, настільки ж чужі для еволюції, як і наш Творець чужий для нас. Одна чиста функція корисності розсіялася на тисячі відтінків бажання.

Як таке могло статися? Передусім завдяки абсолютній дурості еволюції. Крім того, *перші* білкові комп'ютери були далекі від *загальних*, як і їхній творець, так що вони могли вираховувати користь лише з короткострокових бажань.

Наостанок, питати, чому еволюція не зробила з людей максимізаторів інклюзивної пристосованості, — все одно що питати, чому еволюція не підкинула людям рибосому, давши самим склепати свою біохімію. Еволюція не вміє реорганізовувати код аж так швидко — ось чому. Але, мабуть, через мільярди років подальшого природного добору *закріпиться* і ця адаптація, якщо інтелект не проявить розумності і дозволить придуркуватому богові далі панувати.

У книзі «Порошинка в оці Бога» (The Mote in God's Eye) Ларрі Нівен і Джеррі Пурнелл зображують розумний вид, що надто довго лишався біологічним, внаслідок чого зрештою потрапив у рабство еволюції, а його організми поступово перетворилися на максимізаторів пристосованості, одержимих перевідтворенням себе. На щастя, нас не спіткала така доля. Не на Землі. Принаймні поки що.

Людям до смаку цукор і жир. Люди люблять своїх синів і доньок. Ми постійно у пошуках соціального статусу і сексу. Ми співаємо, танцюємо і граємося. Ми вчимося, бо нам подобається вчитися.

Тисячі звабливих смаків, що колись посилювалися через кореляцію з репродуктивною пристосованістю, тепер цінуються незалежно від того, чи підвищують вони відтворення. Секс з контрацептивами, шоколад, музика давно померлого Баха на CD-диску.

І тепер, коли ми дізнаємося про еволюцію, думаємо про себе: «Цілими днями бути одержимим інклюзивною пристосованістю? Що ж у *цьому* класного?».

Єдина мета сліпого дурнуватого бога розпалася на тисячі відтінків бажання. Думаю, у цьому немає нічого поганого, хоч це й пишу я — людина. Інакше що ще нам робити з майбутнім? Що ми могли б зробити з мільярдом галактик у зоряному небі? Заповнити їх максимально ефективними реплікаторами? Чи мають наші нащадки свідомо обрати одержимість максимізацією власної інклюзивної пристосованості, розглядаючи все інше лише як засоби досягнення цієї кінцевої мети?

Бути втіленням тисяч відтінків бажання не завжди класно, але хоча б не *нудно*. Десь по дорозі ми набули смаку до новизни, складності, гармонії і виклику — усі вони судять сліпого бога і виносять вирок: його мономанія естетично неприйнятна.

Безумовно, ці самі смаки ми успадкували від убивці сліпого придуркуватого бога. То й що з того?

Крихкі цілі

Віра в інтелект

Позаяк я не впевнений у ходах Каспарова, який емпіричний вміст мого переконання «Гаррі Каспаров — надзвичайно розумний шахіст»? Якого безпосереднього досвіду я маю очікувати, тримаючись цього переконання? Чи не маскує воно моє цілковите невігластво під щось витонченіше?

Поставлю дилему руба. Припустімо: Каспаров грає проти пересічного ґросмейстера містера Джі, який не змагається за звання чемпіона світу. Я дуже погано відрізняю ці рівні шахової майстерності. Коли я намагаюся вгадати хід Каспарова або наступний хід містера Джі, щонайбільше я можу спробувати вгадати «найкращий шаховий хід», спираючись на свої мізерні знання з шахів. У такому разі я робив би практично ідентичне передбачення що для Каспарова, що для Джі у будь-якій шаховій позиції. Тож який емпіричний вміст мого переконання «Гаррі Каспаров — *кращий* шахіст за містера Джі»?

Емпіричний вміст мого переконання — це тестовне, фальсифіковне передбачення того, що *кінцева* шахова позиція у класі шахових позицій виявиться виграшною для Каспарова, а не завершиться нічиєю чи перемогою містера Джі. (Законним ходом вважаю і відмову від гри, якщо вона веде до позиції, що розцінюється як поразка.) Міра, в якій я вважаю Каспарова «кращим шахістом», виражається обсягом ймовірнісної маси, яку я зосереджую у кластері результатів «перемога Каспарова» порівняно з кластером «нічия» і «перемога містера Джі». Ці кластери надзвичайно розмиті у тому сенсі, що охоплюють величезні простори можливих шахових позицій, хоч і «перемога Каспарова» — більш конкретна гіпотеза за максимальну ентропію, адже її, безперечно, можна спростувати великою множиною шахових позицій.

Результат гри Каспарова передбачуваний, бо я знаю і розумію його цілі. У межах шахівниці я знаю мотивації Каспарова, його критерій успіху, функцію корисності та ціль як

оптимізаційний процес. Я знаю, у який бік Каспаров намагається, *зрештою*, скерувати майбутнє, й очікую, що він достатньо умілий, щоб дійти до нього, навіть не маючи очікувань щодо того, *як* він це зробить.

Уявімо, що я приїхав у незнайоме місто, і тамтешній друг зголосився підвести мене до аеропорту. Я не орієнтуюся у районі. Щоразу, під'їжджаючи до перехрестя, я не знаю, чи повернe він ліворуч, праворуч чи поїде прямо. Не можу передбачити його наступну дію навіть на кожному окремому перехресті — що вже говорити про передбачення повної послідовності дій наперед.

І все ж я можу передбачити *результат* непередбачуваних дій мого друга: зрештою, ми прибудемо до аеропорту. Навіть якби товариш жив в іншому районі — а отже, обирав би геть іншу послідовність поворотів, — я з тією самою впевненістю очікував би нашого прибуття до аеропорту. Ба більше, я передбачаю це задовго до того, як ми навіть сядемо в автівку. Вже зовсім скоро відлітає мій рейс, і я б не марнував часу на це, якби не був упевнений, що за всієї непередбачуваності маршруту машина таки доправить нас до аеропорту.

Хіба з наукового погляду це не приголомшлива ситуація? Я можу передбачити результат процесу, а його *проміжні етапи* — вже зась.

Як таке взагалі можливо? Зазвичай передбачення будують інакше: зображують у думці теперішній стан і прокручують його у майбутнє. Якщо вам потрібна *точна* модель Сонячної системи, слід враховувати планетарні збурення, почати з моделі всіх основних тіл і далі просувати її вперед у часі — крок за кроком.

Іноді простіші задачі мають розв'язок замкненого вигляду: обчислення стану в момент часу T потребує однакових зусиль незалежно від величини T. Монетка лежить на столі, і щохвилини її перевертають. Процес починається з герба. Якою стороною монетка лежатиме за сто хвилин? Ясно, що ви не уявляли собі сто послідовних обертів. Натомість ви скористалися

розв'язком замкненого вигляду, який дозволив передбачити результат, а за потреби — *і* будь-який з проміжних етапів.

Та коли мій друг підвозить мене до аеропорту, я передбачаю результат за допомогою дивної моделі, що не розкриває *жодного* проміжного етапу. Ба більше, вона навіть не вимагає ввести початкові умови — мені не потрібно знати точку відправлення!

Натомість мені потрібно дещо знати про мого друга. Маю знати, що він хоче, щоб я встиг на рейс. Маю бути впевненим, що мій друг вміє планувати та вчасно довезе мене до аеропорту (якщо захоче). Такими є властивості *його* вихідного стану. Саме вони дозволяють мені передбачити кінцеве призначення, чого не скажеш про проміжні повороти.

Крім того, я маю бути впевненим у тому, що як водій він добре знає місто. Це можна розглядати як відношення між моїм другом і містом — а отже, як властивість обох. Утім, ця властивість надзвичайно *абстрактна* і не потребує жодного *конкретного* знання ні про місто, ні про те, що саме мій друг про нього знає.

Це один із поглядів на предмет дослідження, якому я присвятив життя, — на ці *разючі ситуації*, що ставлять нас у настільки дивну епістемічну позицію. У певному сенсі на мою роботу можна дивитися як на розкриття точної форми того дивного абстрактного знання, яким ми володіємо; й ось так, не знаючи дій, ми обґрунтовано можемо передбачити наслідок.

«Інтелект» — завузький термін для вичерпного опису таких дивовижних ситуацій. Як на мене, краще підходить словосполучення «оптимізаційний процес». Схожа ситуація виникає, наприклад, і при вивченні біологічного природного добору: ми не можемо передбачити точну форму наступного спостережуваного виду.

Моя ж спеціальність — такий собі оптимізаційний процес під назвою «інтелект», а ще вужче — певний вид інтелекту під назвою «Дружній штучний інтелект», з якого, сподіваюся, мені вдасться здобути особливо точне абстрактне знання.

Люди в кумедних костюмах



Уже невідь-скільки разів люди літали у космос у пошуках життя на далеких зорях, але щоразу виявлялося, що ті позаземці — зовсім як люди: тільки в кумедних костюмах, трохи підфарбовані гримом, у латексі, або ж звичайні бежеві кавказці.

Аж не віриться: обриси людського тіла виявляються настільки універсальними у всесвіті, що всі інші позаземні види відрізняються лише кількома незначними модифікаціями.

Яке можливе пояснення цьому дивовижному явищу? Звісно ж, конвергентна еволюція! І хоча ці позаземні форми життя розвивалися на інших планетах, цілковито незалежно від земного життя, усі вони прийшли до однієї точки розвитку.

Нехай вас не вводить в оману те, що кенгуру (ссавець) менш подібне до нас за шимпанзе (примат), або те, що жаба — в амфібій, як і в нас, чотири кінцівки — схожа на нас менше, ніж кенгуру. Нехай вас не вводить в оману і приголомшливе розмаїття комах, що відокремилися від нас ще раніше за жаб; нехай вас не

збивають з пантелику ані шість кінцівок у комах, ані їхній зовнішній скелет, ані інша система зору, ані навіть відмінні способи розмноження.

Певно, ви думаєте, що справжній позаземний вид відрізнявся б від нас ще більше, ніж ми від комах: позаземці не виходили б із ДНК і не складалися б зі складених вуглеводних ланцюжків, внутрішньо скріплених вандерваальсівським зв'язком (білками).

Як я вже писав, нехай усе це не вводить вас в оману. Щоб позаземний вид еволюціонував до *інтелекту*, у нього мають бути дві ноги з одним коліном кожна, прикріплені до вертикального тулуба, і пересуватися він має приблизно так само, як ми. Бачте, жоден *інтелект* не обійдеться без рук, а для цього потрібно переформатувати одну пару кінцівок; якщо ця істота не має чотирьох кінцівок, вона, звісно, не розвине здатності бігати і ходити прямо без допомоги рук.

Щоб позаземний вид еволюціонував до *інтелекту*, йому потрібен бінокулярний зір для точної взаємодії з реальністю — отже, конче необхідні саме два ока. Ці очі мають розташовуватися на голові, що сидить на тулубі, а не десь-інде. Позаземець комунікує виключно шляхом перекладу думок в акустичні вібрації, а для цього йому потрібні вуха, губи й горло. А як вже є вуха, очі й губи, то куди ж без носа! Статевий добір у жодному разі не залишить без носа. От вам захотілося б сходитися з кимось без обличчя? Отож. Схожа логіка пояснює й те, чому жіночі особини цього виду без винятку привабливі: потворні інопланетяни просто не мали б достатнього репродуктивного успіху. А чому позаземці розмовляють англійською? Бо якби розмовляли якоюсь абракадаброю, то дідька лисого збудували б повноцінну цивілізацію.

...або суто заради *альтернативної* гіпотези *можемо* припустити, що людей у кумедних костюмах використовувати *куди простіше*.

Але справа навіть не в формі — вона в розумі. Термін «люди в кумедних костюмах» добре відомі у фанатських спільнотах

наукової фантастики. І він *не* позначає прямоходячу істоту з чотирма кінцівками. Кутаста істота з чистого кришталю являє собою «людину в кумедному костюмі», якщо вона *мислить* точнісінько як людина, особливо людина з англомовного світу кінця XX — початку XXI століття.

Я не дивлюся багато старих фільмів. Кілька років тому, переглядаючи стрічку «Психо» (1960), я гостро відчув культурну прірву між Америкою, яку я побачив на екрані, та Америкою, яку знав. Оці персонажі в сорочках із ґудзиками здалися мені значно чужішими за добру частину так званих позаземців, яких я бачу по телевізору чи на срібному екрані.

Щоб змалювати культуру, не схожу на власну, варто навчитися бачити свою як *окремий випадок*, а не норму — не як точку відліку, від якої мають відштовхуватися всі інші культури. У цьому може допомогти історія, хоча й вона — лише чорні літерки на білому папері, а не ваш живий досвід. Думаю, від року проживання у Китаї, Дубаї чи навіть у племені кунґ було б більше користі... і навіть цього я не зробив через мою зайнятість. Часом мені думається, скількох речей я міг не помічати — навіть не в іншій культурі, а у своїй.

А розглядати *все людство* як окремий випадок ще складніше.

У кожній відомій культурі люди, вочевидь, переживають радість, сум, страх, відразу, злість і здивування. У кожній відомій культурі ці емоції відображаються однаковими виразами обличчя. Тож наступного разу, коли побачите «позаземця» — чи, зрештою, «штучний інтелект», — закладаюся, що при гніві (а він таки гніватиметься) на його обличчі з'являться риси, характерні для будь-якої людини у стані злості.

Ми, люди, дуже подібні між собою тим, що у нас під черепами, — так уже влаштовані види зі статевим розмноженням: не може бути так, щоб кожен мав набір радикально різних складних адаптацій — вони просто не змогли б зібратися в єдину систему. (Чи розмножуються позаземці статевим шляхом, як люди та багато комах? Чи мають вони невеликі спільні фрагменти

генетичного матеріалу, як бактерії? Чи формують вони колонії, як гриби? Чи діє серед них закон психологічної єдності?)

Вашим предкам доводилося *маніпулювати — деталізовано* моделювати, а не просто приручати чи піймати у сіті — розумами, які працювали більш-менш подібно до їхніх власних. Так ми еволюціонували до навички передбачати Інші Розуми, навчилися ставити *себе* на їхнє місце, питати, що ми зробили б у їхній ситуації. Адже все, що потребувало передбачення, було подібним до того, хто передбачає.

«Що? — скажете ви. — Я не припускаю, що інші люди подібні до мене! Коли я сумую, вони можуть злитися! Вони вірять не в те, у що я; їхні особистості відмінні від моєї!». Подивіться з цієї сторони: людський мозок — *надзвичайно* складна фізична система. Ви не моделюєте його нейрон за нейроном чи атом за атомом. Якби вам трапилася фізична система складністю з людський мозок, *не* схожа на вашу, знадобилося б кілька поколінь науковців, щоб розплутати її. Ви *не* розумієте, як функціонує людський мозок у загальному, абстрактному сенсі; ви не можете створити мозок чи навіть комп'ютерну модель, що передбачає активність інших мізків так, як це робите *ви*.

Годі шукати якоїсь іншої причини того, що ви *взагалі* здатні осягнути щось настільки фізично складне і заледве вивчене, як мозок іншої людини, крім тієї, що ви налаштовуєте свій мозок на імітацію іншого. Ви ставите себе на місце іншого, хоч, певно, і не співпереживаєте йому. Ви накладаєте тінь злості іншого розуму і тінь його переконань на свій мозок. У вашому розумі може ніколи й не з'явитися формулювання «Що б я зробив у цій ситуації?» — аж раптом та маленька тінь іншого розуму оживає у вашому, чим запускає той самий складний механізм, що існує в голові іншої людини. Так його механізм синхронізується із вашим. І вам не втямки, як усе це відбувається. Можливо, ви зараз не злитеся, але ви знаєте: якби *ви* злилися на самого себе і думали про себе як про останнього виродка, то *ви* спробували б прибити себе...

Ось так «емпатичне висновування» (назву це так) більшменш працює у людей.

А як щодо розумів з *інакшими* емоціями, яких ви ніколи не переживали, або розумів, які ніколи не відчували тих емоцій, з якими ви добре знайомі? Через прикладання свого мозку до іншого такого вже не змоделювати. Я можу попросити вас уявити позаземця, котрий виріс у всесвіті, де простір чотиривимірний (а не тривимірний), але ж у вас не вийде переналаштувати свою зорову кору, щоб бачити так, як той позаземець. Я можу постаратися написати історію про позаземців з іншими емоціями, але що у вас, що у мене однаково не вийде пережити їх.

Уявіть, як позаземець дивиться постановки братів Марксів[1] і поняття зеленого не має, що відбувається на екрані або чому ви активно шукаєте подібних вражень, бо той прибулець і близько не сприймав чогось схожого на почуття гумору. А як вже почали співчувати їм через те, що вони втратили, не забувайте: *ви* ніколи не *рогатіли*.

І тут, я певен, декілька читачів візьмуться фантазувати, чому еволюція неминуче породжує інтелект з почуттям гумору, навіть якщо вона взагалі його породжує. А може, позаземці і мають почуття гумору, а це у вас несмішні жарти? Це точно як в іншій країні намагатися говорити англійською дуже гучно і дуже повільно — ніби в теорії іноземці мають внутрішнього духа, що вловлює *значення*, притаманні вашим словам, якщо тільки ви вимовляєте їх досить голосно, аби подолати якийсь дивний бар'єр, що стоїть на заваді вашій універсально зрозумілій англійській.

Не без того, щоб цінувати таке прекрасне і варте явище як сміх, навіть якщо він виявиться не універсалізовним, навіть якщо він не притаманний усім можливим розумам. Нехай він став би нашою *особливою* складовою дару, який ми передаємо завтрашньому дню. Це теж може мати значення. І краще б воно мало, адже універсалізовність — одне з тих метаетичних понять,

1. https://uk.wikipedia.org/wiki/Брати_Маркс

що я не можу для вас врятувати. Щось може бути універсалізовним серед людей, але не серед усіх можливих розумів.

Без протиставлення ми не мислимо себе як людей. Режисери, які зображували прибульців викрадачами дівчат у розірваному одязі, припустилися цієї помилки не через осмислення імовірної еволюції позаземних розумів. Просто у них склалося враження, що дівчина у роздертому одязі викликає хіть, причому ця властивість стосується як дівчини, так і її одягу, хоча до позаземців жодного стосунку вона не має. Англійські слова *містять* значення, жарти *є* кумедними. Який стосунок попереднє речення має до позаземців?

Антропоморфізм так глибоко вкорінився у наших розумах, що годі викорчувати його самим лише зусиллям волі чи рішучою заявою: «Відтепер я присягаюся більше не думати як людина!». Людськість — що повітря, яким ми дихаємо; *полотно*, на якому накидаємо ескізи. Навіть якщо уявити, як слизьке чудовисько спаровується з подібними йому, все ж важко *позбутися* думки, що воно віддасть перевагу дівчині в розірваному одязі як бажанішій і спокусливішій здобичі. Складно уявити, що воно *не* скаже: «А знаєте: це ж я спаровувався зі слизькими чудовиськами, допоки не побачив землянок. До біса їх — землянки куди кращі».

А як щодо ро́зумів без емоційних систем, у яких навіть немає аналогів емоцій? Ні, не поспішайте пояснювати, чому інтелектуальний розум, здатний будувати складні механізми, неминуче розвине стани, аналогічні до емоцій. Натомість поцікавтеся еволюційною біологією: природний добір створює складні механізми й при цьому не має емоцій. А тепер *представляю* вам справжнього чужого — оптимізаційний процес, що функціонує *зовсім* не так, як ви.

Значна частина прогресу в біології від 1960-х років полягала у спробі накласти мораторій на антропоморфізацію еволюції. То була головна академічна сварка, і я не певен, чи взяв би глузд гору, якби не переконливі експериментальні свідчення, підкріплені чіткою математикою. Змусити людей перестати ставити

себе на місце Іншого — довга і важка робота. У галузі ШІ я борюся за це роками.

У науковій фантастиці лакмусом таланту автора є його спроможність виписати справжніх чужих. (І не просто якихось незбагненних позаземців, зручних авторові тим, що їхньою таємничістю можна виправдати будь-який сюжетний поворот.) Джек Венс був серед корифеїв цього мистецтва. Якщо Венс вписував у свій твір *людей* іншої культури, вони були більш чужими, ніж більшість «чужопланетян». (Кажете, ніколи не читали його? Раджу почати з роману «Місто Чаш» (City of the Chasch).) Куди ж без «Порошинки в оці Бога» Ні́вена і Пурне́лля.

І навпаки — пригадую, як читав, що один письменник (здається, О́рсон Скотт Кард) нарікав на найгірший епізод «Зоряного шляху» за всю історію серіалу. У ньому йшлося про паралельну еволюцію, яка породила позаземців, що не лише були схожі на людей і розмовляли англійською, а й незалежно, слово в слово, переписали преамбулу до Конституції США.

Якесь глибочезне дно уяви. І не думайте, що справа лише у НФ чи ШІ. Нездатність уявити іншопланетянина обертається нездатністю побачити *самого себе* й усвідомити власну особливість. Хто ж тоді розпізнає людину, замасковану серед людей?

Оптимізація і вибух інтелекту

Оптимізаційний процес — одна з тем, яких я ще не торкався. У загальних рисах ця ідея полягає в тому, що силою розуму в широкому просторі пошуку — чи то в просторі можливих сценаріїв (планування), чи то в просторі можливих проєктів (винахідництво) — можна знаходити способи досягнення конкретних цілей.

Нехай у вас є автомобіль, а серед ваших цілей — мандри. А тепер уявімо, що ви берете всі запчастини або атоми автомобіля і безладно перемішуєте їх. Дуже малоймовірно, що в такий спосіб колись з'явиться засіб для подорожей — навіть щось схоже на візок на колесах, не кажучи вже про засіб пересування, який посідав би у вашому переліку уподобань таке саме високе місце, як і власне автомобіль. Отже, з огляду на порядок ваших уподобань, автомобіль — надзвичайно *неймовірний* артефакт. Сила ж оптимізаційного процесу у тому, що він здатен породжувати такий вид неймовірності.

Інтелект і природний добір можна розглядати як особливі випадки *оптимізації*: обидва є процесами, котрі у широкому просторі пошуку досягли крихітних цілей, визначених непрямими вподобаннями. Природний добір надає перевагу ефективнішим реплікаторам. Людський інтелект має складніші вподобання. Ні еволюція, ні люди не мають послідовних функцій корисності, тому їх розуміють як «оптимізаційні процеси» лише наближено. Ви намагаєтеся розібратися у тому, *яка саме робота виконується*, а не стверджувати, що еволюція чи людина виконує її *досконало*.

Такою я бачу історію життя та інтелекту — історію неймовірно вдалих проєктів, створених оптимізаційними процесами. Тут «неймовірно» означає неймовірний відносно випадкового добору з простору можливих проєктів, а не неймовірність в абсолютному сенсі: якщо десь поруч діє оптимізаційний процес, «неймовірно» вдалі проєкти стають імовірними.

При огляді історії оптимізації на Землі аж до наших днів насамперед варто провести поняттєву межу між метарівнем та рівнем об'єктів — між *структурою оптимізації* та *об'єктом оптимізації*.

Якщо з біології вилучити гомінідів, то на рівні об'єктів залишаться динозаври, метелики й коти, а на метарівні — статева рекомбінація та природний добір асексуальних популяцій. І ви побачите, що рівень об'єктів куди складніший за метарівень. Хай природний добір — непросте явище, де без математики нікуди, але якщо подивитися на анатомію цілого кота, то процеси всередині кота набагато складніші за «мутувати, рекомбінуватися, розмножуватися».

І в цьому немає нічого дивного. Адже природний добір — це *випадковий* оптимізаційний процес, котрий, по суті, почався одного дня десь у припливно-відпливному басейні. А кіт — *об'єкт* і результат мільйонів і мільярдів років еволюції.

У кота є мозок, який навчається протягом усього життя, але після його смерті вся зібрана інформація зникає, не лишаючи сліду. Отже, сукупний оптимізаційний вплив котячих мізків на світ відносно незначний.

Візьмемо мозок бджоли чи бобра. Бджоли будують вулики, бобри — греблі, але вони не з'ясовують, як побудувати їх з нуля. Бобер не зможе збудувати вулик, а бджола — греблю.

Отже, мізки тварин — донедавна — не були головними гравцями у планетарній грі оптимізації: вони були *фігурами* — не *гравцями*. Порівняно з еволюцією мізкам бракувало загальної сили оптимізації (вони не могли створювати дивовижне розмаїття артефактів, породжених еволюцією) і кумулятивної сили оптимізації (їхні продукти не накопичували складність із часом). Цю тему я більше розкриваю в есеї «Білкове підкріплення та консеквенціалізм ДНК[1]».

Зовсім недавно у деяких тварин почали з'являтися ознаки загальної сили оптимізації (вони змогли створювати напрочуд

1. (https://lesswrong.com/lw/l2/protein_reinforcement_and_dna_consequentialism/)

широкий спектр артефактів за часовий проміжок надто короткий, щоб еволюція могла суттєво на нього вплинути) і кумулятивної сили оптимізації (артефакти ставали дедалі складнішими, а навички передавалися через мову і письмо).

Будь-які зміни природного добору займають сотні поколінь і мільйони років, щоб з'явилися нові складні проєкти. Програмісти за день можуть створити новий механізм із сотнею взаємозалежних компонентів. І це не дивно, бо природний добір — *випадковий* оптимізаційний процес, що, по суті, почався одного дня, тоді як люди — *оптимізовані* оптимізатори, сконструйовані природним добором протягом мільйонів років.

Диво еволюції не в тому, наскільки ефективно вона працює, а в тому, що вона взагалі функціонує без попередньої оптимізації. Ось так оптимізація і започаткувала своє існування — з украй неефективного випадкового оптимізаційного процесу, як і слід було очікувати. І мова не про перший випадковий ген-реплікатор, а про перший випадковий процес природного добору. Розрізняйте рівень об'єкта і метарівень!

Від самого зародження оптимізації у всесвіті існувала певна структурна подібність як у природному доборі, так і в людському інтелекті . . .

Природний добір *відбирає гени*, але загалом гени не розвертаються і не беруться оптимізовувати сам природний добір. Поява статевої рекомбінації — виняток із цього правила, як і поява клітин і ДНК. Потужний вплив і водночас *рідкісність* таких подій видно й сьогодні: еволюційні біологи структурують цілі історії життя на Землі довкола появи клітин і ДНК.

Та якщо зробити крок назад і подивитися на все з людської перспективи — мислити як програміст, — тоді можна побачити, що добір *все ще* не такий уже й складний. Спробуємо скомбінувати різні гени? Спробуємо відокремити інформаційне сховище від рухомого обладнання? Спробуємо випадково рекомбінувати групи генів? До таких ідей у масштабі всієї системи будь-який кмітливий хакер міг би дійти за перші десять хвилин роздумів про архітектуру.

Через те, що природний добір мав настільки неефективний початок (як цілковито випадковий процес), ці крихітні метарівневі покращення, що надходять від генів-реплікаторів — значно простіші за анатомічну будову кота, — стали ключовими віхами розвитку земного життя.

І навіть *після* всього цього природний добір *усе ще* лишається сліпим придуркуватим богом: попри всі клітини та статеве розмноження, генофонди можуть доеволюціонуватися до вимирання.

Природний добір справді підживлює себе сам у тому сенсі, що кожна нова адаптація відкриває нові шляхи для подальших — але лише на рівні об'єкта. Генофонд підживлюється своєю складністю, але лише завдяки захищеному інтерпретатору природного добору, що працює у фоновому режимі, не переписує сам себе і не змінюється внаслідок еволюції видів.

Схожим чином люди винаходять науки й технології, але ще не почали переписувати захищену структуру самого мозку. Як і в перших аграріїв, у нас є префронтальна кора, скронева частка і мозочок. Ми ще не почали генетично модифікувати власний вид. На рівні об'єктів наука підживлює науку: кожне нове відкриття прокладає дорогу новому. Але все це можливе завдяки захищеному інтерпретатору — людському мозку, що працює у фоновому режимі, залишаючись недоторканим.

У нас є і метарівневі винаходи — як-от наука, — котрі намагаються настановляти людину на те, як мислити. Але людина, яка першою сформулювала теорему Баєса, баєсіанцем так і не стала: він не зміг переписати себе, бо не мав ані відповідних знань, ані такої можливості. Наші великі інновації у мистецтві мислення — письмо і наука — визначають хід людської історії, хоч і за складністю до мозку їм ще далеко, а їхній вплив на мозок порівняно неглибокий.

Нинішній стан[2] школи раціональності ще не дає інструментів перетворити випадкового смертного на Альберта

2. https://rationality.org/

Айнштайна — і це добре показує силу кількох незначних генетичних особливостей мозку порівняно з усіма книжками із самодопомоги XX століття.

Оскільки мозок тихо працює собі на задньому плані, люди схильні недооцінювати його внесок і сприймати його як щось самоочевидне — ніби правило «перевіряйте ідеї експериментально» або критерій значущості $p < 0{,}05$ дорівнює за важливістю самому мозку. Спробуйте-но пояснити шимпанзе, що свої ідеї треба перевіряти експериментально, і подивіться, що з того вийде.

А тепер . . . деякі з нас *хочуть* створити розумний проєкт інтелекту, що зможе розумно перепроєктовувати самого себе аж до рівня машинного коду.

Спочатку машинний код, а згодом і закони фізики стануть своєрідним захищеним рівнем. Але той «захищений рівень» уже не міститиме *динаміку оптимізації*; захищені рівні не будуть структурувати роботу. Людський мозок сам по собі досить незле оптимізується і сам по собі помиляється, хай там що ви намагаєтеся розказати йому у школі. Але цей *повністю охоплювальний рекурсивний оптимізатор* не мав би захищеного рівня, який би *оптимізовував*. Вся структура оптимізації підлягала б самій оптимізації.

І ця кардинальна зміна розриває зв'язок з усім минулим — від першого реплікатора, бо вона порушує ідею захищеного метарівня.

Донині історія Землі була історією оптимізаторів, які працювали зі сталою швидкістю і створювали постійний оптимізаційний тиск. Вони породжували продукти оптимізації — *не* з постійною швидкістю, а з прискоренням, оскільки інновації на рівні об'єктів відкривали шлях до інших. Але це прискорення можливе завдяки захищеному метарівню, на якому й відбувається власне оптимізація. Це нагадує пошук, який перескакує між окремими ділянками простору можливих рішень: якщо знайдено хорошу ділянку, поруч зазвичай є ще кращі. Проте сам механізм пошуку при цьому не змінюється. *Іноді* кілька

незначних змін досягають метарівня — як секс або наука, — і тоді починається нова віха в історії оптимізації, де все рухається швидше.

Уявіть економіку без інвестицій, університет без мови, технологію без інструментів для виготовлення інструментів. Раз на сто мільйонів років чи на кілька століть хтось винаходить молоток.

Саме так досі відбувалася оптимізація на Землі.

Коли я дивлюся на історію Землі, я бачу не графік «час — оптимізація», а співвідношення між *оптимізаційною потужністю* на вході та *оптимізованим продуктом* на виході. Дотепер завдяки існуванню майже захищених метарівнів стало можливим розділити історію оптимізації на віхи, а *всередині* кожної віхи — зобразити часовий графік кумулятивної *оптимізації на рівні об'єктів*, оскільки захищений рівень працює у фоновому режимі, не змінюючись всередині неї.

Що відбувається, коли ви створюєте рекурсивно самовдосконалювальний ШІ, здатний оптимізувати всю внутрішню архітектуру? Тоді ви берете графік «оптимізація на вході — оптимізований результат на виході» і згортаєте його на себе. Метафорично.

Якщо ШІ слабкий, нічого не відбувається: він не настільки потужний, щоб покращити себе, як і шимпанзе, що не може перепрошити свій мозок.

Якщо ШІ досить потужний, щоб переписувати самого себе так, аби збільшувати свою здатність до подальших удосконалень, і якщо це сягає аж до повного розуміння ним власного вихідного коду та власного дизайну як оптимізатора . . . тоді навіть якщо графік «оптимізаційна потужність на вході — оптимізований продукт на виході» виглядатиме майже так само, графік оптимізації *в часі* буде разюче відрізнятися від усієї попередньої історії Землі.

Часто кажуть: «Але що як для лінійного вдосконалення знадобиться експоненційно більше зусиль для самопереписування?». На це питання є очевидна відповідь: «Природний добір

впливав приблизно з однаковою оптимізаційною силою на лінію гомінідів у процесі розвитку людини, що, здається, не вимагало експоненційно більше часу для кожного лінійного кроку вдосконалення».

Та це все ще просте аналогічне мислення. Те, як мислить повний загальний штучний інтелект про природу оптимізації, його власні дослідження ШІ, переписування власного вихідного коду, *насправді* не те саме, що графік історії Землі, згорнений на себе. Це інший звір. Такі аналогії *у кращому разі* придатні для якісних передбачень, і навіть тоді у мене є чимало інших переконань, які я ще не пояснив, але які підказують мені, *які саме* аналогії проводити, і так далі.

Але якщо вам кортить знати, чому я не схильний продовжувати графік біологічного й економічного зростання *в часі* — у майбутнє і за горизонт ШІ, що мислить на транзисторних швидкостях, винаходить самовідтворювані молекулярні нанофабрики та *вдосконалює власний вихідний код*, — то ось причина: ви будуєте неправильний графік. Правильний графік має показувати оптимізаційну потужність на вході проти оптимізованого продукту на виході, а не оптимізований продукт залежно від часу.

Духи у машині

Люди чують про Дружній ШІ й кажуть (ось одна з трьох найпоширеніших перших реакцій):

«О, то ви можете сказати ШІ бути дружнім, але якщо він може переписати власний вихідний код, то просто прибере усі обмеження, які Ви на нього наклали».

Звідки береться *таке* рішення?

Воно впало на нас з неба чи є наслідком закономірного ланцюга причин, котрий почався з вихідного коду у його початковому вигляді? Чи є ШІ автором і джерелом власної свободи волі?

Дружній ШІ — це не егоїстичний ШІ, загнузданий спеціальним додатковим модулем совісті, що стримує природні імпульси ШІ і розказує, що йому робити. Ви одразу кодуєте совість, котра і є власне штучним інтелектом. Якщо у вас є програма, що обчислює, яке рішення має ухвалити ШІ, — *справу зроблено*. Далі копати не треба.

На цьому моменті дозволю собі згадати декілька показових випадків з сайту «Комп'ютерні натупи» (Computer Stupidities) і підтему «Програмування». (Посилань не даю — залипнути там можна надовго; хто захоче, той нагуглить.)

* * *

Якось я підробляв репетитором для студентів коледжу, які проходили курс з комп'ютерного програмування. Деякі мої учні не розуміли, що комп'ютери не наділені свідомістю. Не один у своїй програмі на Паскалі використовував коментарі для детальних пояснень: «Тепер виведи ці літери на екран». Одного я запитав, навіщо він це робить. Відповідь: «А як іще комп'ютер зрозуміє, що я

від нього хочу?». Видно, думали, що як їм не до снаги зрозуміти Паскаль, то й комп'ютер не зможе.

* * *

У коледжі я давав приватні уроки у математичній лабораторії однієї школи. Приходить учень — у нього не працює програма мовою BASIC. Він проходив курс для початківців: його завданням було написати програму для опрацювання рецепта вівсяного печива на певну кількість людей. Я подивився на програму. Виглядала вона якось так:

10 Нагріти духовку до 350 20 Усі інгредієнти поєднати у великій мисці 30 Перемішати до однорідної маси

* * *

Якось студент з перших курсів програмування попросив мене поглянути на його програму і пояснити, чому після такого простого обчислення програма завжди повертала нулі. Я поглянув на програму. Там було усе ясно:

```
begin
read("Число яблук", apples)
read("Число морквин", carrots)
read("Ціна за 1 яблуко", a_price)
read("Ціна за 1 моркву", c_price)
write("Всього за яблука", a_total)
write("Всього за моркву", c_total)
write("Всього", total)
total = a_total + c_total
a_total = apples * a_price
```

```
c_total = carrots * c_price
end
```

Я: Твоя програма не може вивести правильні результати до того, як вони будуть обчислені. Він: Га? Правильне рішення логічне ж. Комп'ютер має правильно перевпорядкувати інструкції.

* * *

Так само інстинктивно можна уявити сценарій «програмування ШІ», що мапиться на схожу людську поведінку: сказати людині, що робити. Ніби «програма» дає розпорядження Ка́сперу, що сидить у машині, проглядає вказівки й вирішує, подобаються вони йому чи ні.

Немає ніякого духу, який проглядає інструкції й вирішує, як їх виконувати. Та́ програма і *є* штучним інтелектом.

Це вам не джин, який виконає будь-яке бажання. І не в усьому покірний раб, який виконає будь-яке ваше бажання так, як ви хочете. Єдиний дух (принаймні під час завантаження) — *це* ваші інструкції.

Саме тому, що ви не можете просто *сказати* штучному інтелекту, що робити, він набагато складніший за людські інстинктивні уявлення про ШІ. Потрібно створити дух з нуля; до того ж ви мусите знати, як дух має побачити те, що очевидне для вас, — інакше дух так і не побачить це. Не можна просто *сказати* духові побачити це. Ви маєте з нуля створити те, що бачите.

Якщо ви не знаєте, як створити щось, що, видно, має такі непередавані елементи, як, скажімо, «ухвалення рішень», тоді не можна просто стиснути плечима і віддати всю роботу свободі волі. У підсумку ви лишитеся і без машини, і без духу.

Програма для гри в шахи створюється не швидким процесором єдиним — для ефективної роботи ШІ — і подальшим

запитом «Зроби хід, який *ти* вважаєш найкращим». Мабуть, ви думаєте, що як з самих програмістів кепські шахісти, то й будь-яка порада, яку вони дадуть електронному надмозку, просто сповільнить дух усередині машини. Але ж, бачте, проблема в тому, що нема ніякого духу.

Немає і простого закляття, яким можна — вуф! — викликати цілого духа у машину. Не можна сказати: «Я прочитав закляття — і дух з'явився. Ось вам причина і наслідок». (І навіть якщо замість «прочитати закляття» ви використаєте поняття «емерджентності» або «складності», воно однаково не спрацює.) Програмі не можна просто дати команду: «Будь вмілим шахістом!». Ви маєте зазирнути всередину таємниці мисленнєвого процесу, що стоїть за грою у шахи, і створити усього духа з нуля.

Немає значення, наскільки нормальним, логічним, «очевидним», «правильним», «самоочевидним» чи «розумним» вам щось здається — воно не відбудеться всередині духу. *Хіба що* воно відбудеться наприкінці причиново-наслідкового ланцюга, що розпочався з інструкцій, які ви дали програмі, разом з якимись причиново-наслідковими залежностями від сенсорних даних, що *ви* їх вбудували в початкові інструкції.

Це не означає, що ви прямо програмуєте кожне рішення. Комп'ютер Deep Blue грав у шахи набагато краще за його програмістів: він робив ходи куди кращі за ті, на які могли його змогли б запрограмувати безпосередньо, але не тому, що програмісти стенули плечима і залишили всю роботу духові. Комп'ютер робив кращі ходи за програмістів... наприкінці причинового ланцюжка, що розпочинався у програмістському коді, і звідти ішов закономірним шляхом. Нічого не відбулося *просто* тому, що якийсь хід був таким очевидним, що дух Deep Blue зі своєю свободою волі вибрав його без втручання коду і закономірних наслідків.

Якщо ви спробуєте зняти з себе відповідальність за обмеження ШІ, ви не отримаєте вільного духа, як то буває зі звільненими рабами. Ви отримаєте лише купу піску, яку ніхто не

очистив до кремнію, не перетворив на центральний процесор і не запрограмував думати.

Гайда! Розкажіть комп'ютерному чипу, що він може робити що схоче! Цікаво: що ж станеться? Нічого. Бо ви не обмежили його до розуміння свободи.

Усе, що треба, — один-єдиний крок, котрий настільки очевидний, логічний, самоочевидний, що ваш розум просто пропускає його. Як наслідок, ви сходите зі шляху програміста штучного інтелекту. А щоб не дати розумові наступати на ці граблі, потрібні зусилля, подібні до тих, що я їх проілюстрував ув есеї «Як вхопити слизьке» (Grasping Slippery Things[1]).

1. https://www.lesswrong.com/posts/HnS6c5Xm9p9sbm4a8/grasping-slippery-things

Штучне додавання

Припу́стимо, що люди *не мали й крихітного уявлення* про те, як виконувати арифметичні операції. Уявімо, що люди не *навчи́лися* рахувати й додавати овець, а розвинули цю навичку *еволюційним* шляхом. І от, користуючись цією вбудованою здатністю, люди гадки не мають, як вона працює, як Арістотель гадки не мав, як зорова кора у його мозку підтримувала його здатність бачити. Арифметика Пеа́но у тому вигляді, в якому ми знаємо її сьогодні, ще не була винайдена. Не без філософів, які працюють над формалізацією числових інтуїцій, але позначення, якими вони користуються, виглядають як

Plus-Of(Seven, Six) = Thirteen

За допомогою них вони формалізують інтуїтивно очевидну істину: коли ви додаєте «сім» до «шести», отримуєте «тринадцять».

У цьому світі кишенькові калькулятори працюють на основі збережених велетенських таблиць пошуку арифметичних тверджень, введених вручну фаховою командою штучних арифметиків, для початкових значень у діапазоні від нуля до ста. Хоча зі своєю прагматичною функцією такі калькулятори справляються без проблем, багато філософів стверджують, що вони лише *симулюють* додавання, а не власне *додають* числа. Жодна машина не вміє *рахувати*, тому перед введенням «тринадцять» у калькулятор людина має порахува́ти тринадцять овець. Калькулятори можуть повертати збережені твердження, але вони не знають, що ті означають: введете «двісті плюс двісті» — калькулятор поверне «Помилка: Перевищено ліміт», хоча відповідь інтуїтивно *очевидна*, якщо ви *знаєте значення* тих слів.

Філософи, певна річ, не настільки наївні, щоб піддатися цим інтуїціям. Числа складають чисту формальну систему: ярлик «тридцять сім» має сенс не через внутрішні властивості

всередині слів, а тому, що він позначає тридцять сім овець у зовнішньому світі. Число наділяється цією позначувальною властивістю у *семантичній мережі* відношень до інших чисел. Тому у комп'ютерних програмах ліспі-токену «тридцять сім» не потрібна *внутрішня* структура — адже він змістовний лише як посилання і відношення, а не якась обчислювальна властивість «тридцяти семи».

Ніхто ще не розробив загального штучного арифметика, хоча, звісно, існує багато спеціалізованих, вузьких штучних арифметиків, що працюють із числами між «двадцять», «тридцять» і так далі. І якщо ви поглянете, як довго йшов прогрес із числами у діапазоні «двісті», стає зрозумілим, що ще не скоро нам світить загальний штучний арифметик. Найкращі експерти галузі прогнозують, що калькулятори зможуть додавати на рівні дванадцятирічної людини щонайменше за сто років.

Однак не всі погоджуються із цим прогнозом або просто із загальноприйнятими[1] уявленнями про штучну арифметику. Часто можна почути наступні твердження:

- «Це проблема формулювання: чому дорівнює "двадцять один плюс" залежить від того, чи далі йде "плюс три" чи "плюс чотири". Якщо ми просто зберемо достатньо арифметичних тверджень, щоб охопити загальновідомі істини, то побачимо справжнє додавання в мережі».
- «Але вам життя не вистачить, щоб руками експертів вносити аж стільки арифметичних тверджень. Нам треба, щоб штучний арифметик міг *вчитися* на розлогій мережі відношень між числами, які люди засвоюють дітьми, коли бачать множини яблук».
- «Ні. Штучний арифметик повинен розуміти природну мову, тому замість того, щоб прямо говорити йому, що двадцять один плюс шістнадцять дорівнює тридцяти семи, він може отримати це знання з Інтернету».

1. https://www.overcomingbias.com/2007/10/outside-the-box.html

- «Чесно, у мене складається враження, що ви переконуєте себе, що можете впоратися з цією задачею. Ніхто з вас до пуття не знає, що таке арифметика, тож ви наосліп перебираєте загальні аргументи: "Нам потрібен ША, що може вивчити X", "Нам потрібен ША, що може дістати X з Інтернету". Я про те, що звучить непогано — ніби ви маєте прогрес. І навіть користь для зв'язків з громадськістю: усі вважають, що вони розуміють запропоноване рішення, але ж насправді ви анітрохи не наближаєтеся до *загального* додавання, хоч і маєте успіхи у спеціалізованому. Певно, ми так і не дістанемося засад арифметики. Для людини непосильна задача».
- «Тому-то нам треба розробляти загального арифметика так, як Природа зробила його — еволюційним шляхом».
- «Низхідні підходи явно не дали результатів у створенні арифметики. Нам потрібен висхідний — якось змусити арифметику *виникати*. Слід усвідомити базову непередбачуваність складних систем».
- «Ви усі помиляєтеся. Минулі спроби створити машинну арифметику були приречені на поразку з самого початку, оскільки вони не мали достатньої обчислювальної потужності. Подивіться на трильйони синапсів у людському мозку — ясно, що ніякий калькулятор не має аж таких великих таблиць пошуку. Нам потрібні калькулятори настільки ж потужні, як і людський мозок. Згідно з законом Мура, таке може статися десь між 4:00 і 4:30 27 квітня 2031 року».
- «Гадаю, машинна арифметика буде розроблена, коли дослідники просканують кожен нейрон цілого людського мозку і введуть у комп'ютер, щоб ми змогли симулювати біологічний ланцюжок, відповідальний за додавання у людей».
- «Не думаю, що нам потрібно чекати аж до можливості просканувати увесь мозок. Нейронні мережі подібні до

людського мозку: можна натренувати їх щось робити й при цьому не розуміти, як це робиться. Ми створимо програми, які займатимуться арифметикою, хоч ми, їхні творці, і не зрозуміємо, як вони це роблять».

- «Але ж теорема Ґьоделя демонструє, що жодна формальна система не зможе покрити базові властивості арифметики. Класична фізика піддається формалізації, тож, щоб додати два та два, мозок, напевно, використовує можливості квантової фізики».
- «Агов, якби людська арифметика була такою простою для відтворення на комп'ютері, ми б не змогли оперувати достатньо великими числами, щоб створити комп'ютери».
- «А ви не чули про уявний експеримент китайського калькулятора* Джона Сьо́рла? Навіть якби у вас був величезний набір правил, які б дозволяли додавати «двадцять один» і «шістнадцять», просто уявіть, що ці слова перекладуть китайською, і ви побачите, що ніякого додавання насправді не відбувається. У цій системі немає дійсних *чисел* — лише ярлики, якими люди позначають числа...»

У цієї притчі більше однієї моралі, і я розповів їх у різних контекстах. Це ілюструє ідею рівнів організації: процесор, до прикладу, може додавати два великі числа, бо ці числа не коти у чорних скриньках, а впорядковані структури з 32 бітів.

Але для цілей подолання упереджень розгляньмо дві моралі:

- Перша — небезпека віри у невідтворювані з власних знань твердження.
- Друга — небезпека танців із бубном довкола базових плутаниць.

Щоб ніхто не звинуватив мене в узагальненні на основі вигаданих свідчень, обидва уроки можна винести і з дійсної історії розвитку штучного інтелекту.

Перша небезпека — штучні арифметики зіштовхнулися із проблемою на об'єктивному рівні: вони функціонували як магнітофони, що відтворювали «знання», створені поза системою, за допомогою процесу, що вони не могли записати його внутрішньо. Людина б сказала ША: «двадцять один плюс шістнадцять дорівнює тридцять сім», ША може записати це речення і видати його назад і навіть зіставити «двадцять один плюс шістнадцять» з «тридцять сім!», проте програми не змогли б самі породити таке знання.

Дуже нагадує, коли фізикові, котрий говорить: «Світло — це хвилі», — вірять на слово, записуючи у пам'ять дивовижні слова, а потім видають їх, коли хтось питає: «З чого складається світло?» — та самі для себе вже не можуть згенерувати це знання. Більше про це розкажу завтра.

Друга мораль — небезпека на метарівні, яка поглинула дослідників штучної арифметики та самовпевнених спостерігачів, — небезпека танців із бубном довкола спантеличливих прогалин у вашому знанні. Сам цей порив робити майже що завгодно, *але не* стиснути зуби, віддатися справі й заповнити бісову прогалину.

Скажете ви: «Це емерджентне явище!» чи «Це незбагненне явище!» — однаково не визнаєте, що необхідно мати якесь базове розуміння, яке можливо здобути, але ви його ще не здобули.

Як ви дізнаєтеся, коли у вас воно з'явиться? І годі шукати іншого шляху, крім як лобом протаранити проблему, вивчити про неї що тільки можна і розглядати її з щонайрізноманітніших кутів, можливо, протягом років. Не на це розрахована академічна система, коли від вас вимагають публікувати щонайменше статтю на місяць. Безумовно, венчурні капіталісти не захочуть вкладатися у це. Ви можете або *ось уже* взятися за створення системи, або відмовитися від цієї ідеї й зайнятися чимось іншим.

Ще раз пробіжіть очима по коментарях вгорі — ніхто з них не прагне розпочати пошуки того, що їм бракує, щоб *розвіяти туман таємничості довкола чисел* і перетворити «двадцять

сім» на щось більше, ніж чорна скринька. Жоден коментатор не усвідомив, що їхні труднощі породили невігластво чи спантеличення у їхніх розумах, а не внутрішню властивість арифметики. Вони і не спробували досягти того стану, в якому об'єкт спантеличення перестав збивати з пантелику.

У своїй книзі «Ймовірнісні міркування в інтелектуальних системах: мережі правдоподібного висновування» Джуда Перл пише, що базове розуміння, що лежить в основі графічних моделей, *незамінне* для завдань, котрі їх вимагають. (Боюся, у кількох словах розказати не вдасться, тому вам доведеться прочитати книгу самим. Я не бачив у Мережі жодних популярних оглядів баєсівських мереж, які б адекватно пояснювали причини, що лежать в основі цих принципів, або важливість саме такого математичного підходу. А от книга Перла — просто знахідка.) Колись існували десятки «немонотонних логік», які незграбно намагалися виразити такі інтуїтивні уявлення, як «Якщо у мене спрацювала сигналізація, імовірно, до мого будинку вдерся грабіжник, але якщо я потім дізнаюся, що поряд зі мною стався невеличкий землетрус, імовірно, ніякого грабіжника не було». Графічна модель надає вам розуміння — математичне пояснення того, чому для цього завдання логіка першого порядку має невідповідні властивості, — і стисло викладає правильне рішення, що одним махом охоплює деталі, зрозумілі для здорового глузду. Поки вам бракує такого розуміння, ви не переставатимете латати логіку то тут, то там, додаватимете все більше правок, щоб примусити її відповідати усьому тому, що здається «очевидно істинним».

Без ключа до проблеми штучної арифметики ви не *дізнаєтеся*, чи можна її розв'язати. Якщо ви не знаєте правил, то не знаєте й того, що для будь-якої дії потрібно їх знати. І так з'являтимуться різного роду розумні ідеї, на позір робочі, як-от: розробити штучного арифметика, що зможе читати природною мовою, чи завантажити мільйони арифметичних тверджень з Інтернету.

І все ж *чомусь* ті розумні ідеї ніколи не працюють. Якось так виходить, що ви «не бачили жодних причин, чому ідея б не спрацювала», тому ще не усвідомлювали перешкод, а не тому, що перешкод не існувало. Схоже на те, як наосліп стріляти у мішень на відстані і після черги сліпих пострілів кричати: «Ви не доведете мені, що я не влучив у ціль!». Та поки ви не знімете пов'язку з очей, ви навіть не берете участі у змаганнях зі стрільби. Коли «ніхто не може довести вам», що ваша цінна ідея *хибна*, це означає, що вам бракує інформації, щоб влучити у маленьку ціль у величезному просторі відповідей. *Поки не переконаєтеся, що ваша ідея робоча, вона не спрацює.*

З історії попередніх ключових відкриттів у галузі ШІ та грандіозних провалів, що передували цим відкриттям, я виводжу важливий і прагматичний урок. *Коли основна проблема — власне невігластво, розумні стратегії для обходу незнання призводять до пострілів у ногу.*

Цінності термінальні та інструментальні

На суто інстинктивному рівні людина, яка планує, поводиться так, ніби розрізняє засоби та цілі. Хочете шоколаду? У супермаркет «Паблікс» якраз завезли шоколад. До нього одна миля їзди на південь по Вашинґтон-авеню. Для цього треба сісти в автівку. Сядете в автівку, якщо відкриєте двері. Відкриєте двері, якщо маєте ключі. Ви кладете ключі у кишеню і готуєтеся поїхати...

...коли раптом чуєте по радіо, що землетрус знищив увесь шоколад у тутешньому «Пабліксі». Що ж, якщо у супермаркеті не лишилося шоколаду, немає сенсу і їхати туди, а також сідати у машину і класти ключі у кишеню. Ви дістаєте ключі, телефонуєте місцевій службі доставки й замовляєте шоколадну піцу. Ммм — смакота.

Нечасто я бачив, щоб люди губилися у планах, які придумували самі. Зазвичай до супермаркету не їдуть, якщо знають, що шоколад скінчився. Крім того, я помітив ще таке: коли люди починають прямо говорити про системи цілей замість своїх бажань, коли вони *згадують* «цілі», а не застосовують їх, вони часто збиваються на манівці. Люди добре вміють планувати на практиці, але не в теорії планування[1]— інакше розробників ШІ було б набагато більше.

За моїми спостереженнями, люди спантеличуються зокрема тоді, коли (радше в абстрактних філософських дискусіях, ніж у побутових розмовах) розглядають різницю між засобами та цілями — більш формально між «інструментальними цінностями» і «термінальними цінностями».

(Для довідки потрібен ще один довгий пост.)

Як на мене, частина проблеми у тому, що людський розум використовує доволі ситуативну систему для відстежування

1. https://www.overcomingbias.com/2007/04/expert_at_versu.html

своїх цілей — вона працює, але не ідеально. В англійській мові не вбудовано чіткого розмежування між засобами і цілями: у реченнях «Я хочу врятувати життя сестрі» і «Я хочу дати пеніцилін сестрі» одне й інше передається дієсловом «хотіти».

Чи можемо ми простою мовою описати різницю, що втрачається?

Перша спроба:

«Інструментальні цінності» бажані виключно за умови, що вони принесуть очікувані наслідки. «Я хочу дати сестрі пеніцилін» не тому, що сестра із пеніциліном — благо саме по собі, а через очікування, що пеніцилін вилікує її від некротичної пневмонії. Якби ви очікували, що, прийнявши пеніцилін, ваша сестра розтане у калюжу, як Лиха Відьма Заходу, ви б щосили берегли її від пеніциліну.

«Термінальні цінності» бажані без обумовлення іншими наслідками: те, що «я хочу врятувати життя сестрі», не передбачає того, що після цього їй введуть пеніцилін.

Перша спроба страждає від очевидних недоліків. Якщо внаслідок порятунку життя моєї сестри Землю поглине чорна діра, я б облишив спроби, трохи поплакав, але так і не дав би їй пеніцилін. Чи означає це, що порятунок сестри зводився не до «термінальної» (чи «самої по собі») цінності, бо теоретично вона все ж обумовлена своїми наслідками? Чи намагаюся я врятувати їй життя *лише* тому, що вірю, що чорна діра не поглине Землю? Здоровий глузд мав би підказати, що тут щось не те.

Забудьте про англійську. Ми можемо створити математичний опис системи рішень, де термінальні (або кінцеві) та інструментальні цінності — окремі та несумісні типи, як числа з цілою частиною та з рухомою комою, мовою програмування — без автоматичного переходу між ними.

Ідеальну баєсівську систему рішень можна створити за допомогою лише чотирьох елементів:

- Стани/результати : тип Outcome[]
    - список можливих станів

- {сестра жива, сестра мертва}

- Дії : тип Action[]
  - список можливих дій
  - {давати пеніцилін, не давати пеніцилін}
- Функція_корисності : тип Outcome -> Utility
  - функція корисності, що відображає кожен можливий стан у міру його корисності
  - (корисність представлена як дійсне число між негативною і позитивною нескінченністю)
  - {сестра жива: 1, сестра мертва: 0}
- Функція_умовної_ймовірності : тип Action -> Outcome -> Probability
  - функція умовної ймовірності, що відображає кожну дію на розподіл ймовірностей станів
  - (імовірність представлена як дійсне число між 0 та 1)
  - {дати пеніцилін: сестра жива, 0,9; сестра мертва, 0,1 ;; не давати пеніцилін: сестра жива, 0,3; сестра мертва, 0,7}

Не хвилюйтеся, якщо не можете зчитати систему типів з першого погляду, — я завжди перекладу англійською. Програмістам опис у вигляді окремих операторів допомагає сформувати чіткі ментальні об'єкти.

І саму систему рішень?

- Очікувана_корисність : Дія A -> (Сума O in Можливі стани: Корисність (O) * Ймовірність(O|A))
  - «Очікувана корисність» дії дорівнює сумі (обчисленій для всіх можливих станів) корисності кожного такого стану, помноженої на умовну ймовірність цього стану за умови здійснення дії.
  - {ОК(дати пеніцилін) = 0,9 ; ОК(не давати пеніцилін) = 0,3}

- Обрати : -> (Argmax A in Дії: Очікувана_Корисність(A))
  - Обрати дію, у якої максимальна «очікувана корисність»
  - {return: дати пеніцилін}

Для кожної дії порахуйте умовну ймовірність усіх можливих станів, потім підсумуйте корисність цих станів, помножену на умовну ймовірність. Оберіть найкращу дію.

Такий математично простий нарис системи рішень. І цей спосіб обчислювати рішення неефективний у реальному світі.

Скажімо, вам потрібна *послідовність* дій для виконання плану. За допомогою формального підходу кожна Дія може позначати цілу послідовність. Але так ми створюємо експоненційно величезний простір, подібно до простору усіх речень, які ви можете надрукувати сотнею літер. Простий приклад: якщо одна з можливих дій на першому кроці — «Відстрілити собі ногу», людина, яка планує ходи, вирішить, що це загалом погана ідея, і виключить *усі* послідовності, що починаються цією дією. І от ми спростили цю структуру у нашому представленні. Тепер у нас немає послідовностей дій — лише прості «дії».

І так — є *кілька незначних ускладнень*. Певна річ, вони є — інакше ми б просто взялися і як створили б справжній ШІ. У цьому сенсі ситуація майже повторюється, як із самою баєсівською теорією ймовірностей.

Але це один із тих разів, коли, *на диво, хорошою ідеєю* буде розглянути абсурдно просту версію перед тим, як додавати будь-які високочолі ускладнення.

Візьмімо філософа, котрий стверджує: «Усі ми кінець кінцем егоїсти; нас переймає лише стан власного розуму. Мати, яка стверджує, що дбає про добробут свого сина, насправді хоче *вірити*, що з її сином усе гаразд, і саме це переконання робить її щасливою. Вона допомагає йому заради свого щастя, а не його». Ви відповідаєте: «А уявімо, що мати жертвує життям, щоб зіштовхнути сина з проїжджої частини, по якій насувається вантажівка. Так вона щасливою не стане — тільки мертвою». Оговтавшись від кількахвилинного заціпеніння, філософ відповість:

«Але вона все одно зробила це, бо *цінувала* цей вибір понад інші — через те відчуття важливості, яке пов'язувала з тим рішенням».

На це ви відповісте: «TYPE ERROR: Не знайдено конструктора для очікувана_корисність -> корисність».

Дозвольте пояснити таку відповідь.

Навіть наш простий формальний підхід ілюструє чітку різницю між *очікуваною корисністю* (чимось таким, що має *дії*) і *корисністю* (чимось таким, що має *результати*). Звісно, ви можете перенести і корисності, й очікувані корисності на множину дійсних чисел. Але це все одно що відображати швидкість вітру і температуру у дійсних числах. Від того вони не стануть тотожними.

Філософ починає з того, що всі ваші корисності повинні задаватися через результати, які зводяться до станів вашого розуму. Якби це було так, ваш інтелект діяв би як *рушій, що спрямовує майбутнє* в ті стани, де ви почувалися б щасливішими. Різниця між майбутніми станами зводилася б лише до вашого стану розуму; вам було б байдуже між будь-якими двома варіантами майбутнього, у яких цей стан однаковий.

І що характерно, ви навряд чи пожертвували б своїм життям для порятунку іншого.

Коли ми заперечуємо, що люди іноді *жертвують* своїм життям, філософ у відповідь переходить до обговорення очікуваних корисностей різних дій: «Відчуття *важливості*, яке вона пов'язувала з тим *рішенням*». Цей різкий стрибок *має* спонукати нас зірватися як ошпарених. Спроба перевести очікувану_корисність у корисність видала б помилку у нашій мові програмування. Англійською ж усе це звучить однаково.

Вибори нашої простої системи рішень мають найвищу очікувану корисність, але це нічого нам не говорить про те, *у якому напрямку цей вибір скеровує майбутнє*. Нічого не каже і про корисності, що їх призначає ухвалювач рішення, чи які реальні наслідки будуть від його рішення. Нічого не каже і про функцію розуму як рушія.

Фізичною причиною фізичної дії є когнітивний стан: у нашого ідеального ухвалювача — очікувана корисність, яку обчислюють, застосовуючи функцію корисності до уявних наслідків. Щоб врятувати синові життя, ви маєте уявити подію, в якій рятуєте йому життя. Але ця уявна ситуація ще не сама подія. Різниця у *лапка́х* — така, як між «снігом» і снігом. Але це і не означає, що *всередині лапо́к* має бути когнітивний стан. Якщо ви обираєте дію, що веде до майбутнього, яке ви описуєте словами «мій син усе ще живий», тоді ви спрацювали як рушій, що спрямував майбутнє в область можливих станів, де ваш син іще живий. Не рушій, що спрямував майбутнє туди, де ви *висловлюєте речення* «мій син усе ще живий». Щоб спрямувати майбутнє *туди*, ваша функція корисності мала б повернути високу корисність, коли їй згодували «"мій син усе ще живий"» (лапки у лапках — уяву про вашу уяву). З рецептів виходять погані пироги, якщо їх подрібнити і додати у тісто.

Тому-то корисно першими розглянути прості системи рішень. Додайте до них достатньо ускладнень — і раніше чіткі відмінності стає важче помітити.

Тепер розгляньмо деякі ускладнення. Очевидно, функція корисності (накладання корисностей на результати) призначена для формалізації того, що я раніше називав «термінальними цінностями» — цінностями, що не залежать від їхніх наслідків. Як щодо випадку, коли порятунок життя вашій сестрі призведе до поглинання Землі чорною дірою? У нашій формальній системі ми виключили таку можливість. Результати не ведуть до результатів — лише дії ведуть до результатів. Ваша сестра одужує від пневмонії, *потім* Землю поглинає чорна діра, і все це стиснуто до єдиного «можливого результату».

Яке ж тоді місце «інструментальних цінностей» у цій простій формальній системі? Вони ж повністю зникли! Бачте, у цій формальній системі дії безпосередньо призводять до результатів, без проміжних подій. Не існує ніякої ідеї кинути камінь, який полетить до яблука, зіб'є його з гілки, і воно впаде на землю. Відповідно до функції умовної ймовірності, що перетворює дію

прямо у розподіл імовірностей за результатами, кидання камінчика — це вже дія, що прямо веде до результату — яблуко лежить на землі.

Щоб *насправді обчислити* функцію умовної ймовірності та окремо розглянути корисність пневмонії сестри і поглинання Землі чорною дірою, нам довелося б представляти мережеву структуру причиновості — те, як події ведуть до інших подій.

І тоді ми почнемо бачити інструментальні цінності. Якби причинова мережа була достатньо постійною, можна було б знайти стан Б, який, як правило, призводив до стану В, безвідносно того, як ви досягали Б. Тоді якби вам забаглося для чогось досягнути В, можна було б скласти ефективний план, як спершу знайти Б, що веде до В, а потім — А, що веде до Б. Так ми й дійшли до явища під назвою «інструментальна цінність»: Б має «інструментальну цінність», бо веде до В. А от саме В може цінуватися уже термінально — як член у функції корисності, що характеризує загальний результат. Або ж В може бути інструментальною цінністю, вузлом, що безпосередньо не оцінювався функцією корисності.

У цій формальній системі інструментальна цінність — це суто костур для ефективного розрахунку планів. Його можна і потрібно відкидати скрізь, де такої постійності не існує.

Нехай є, наприклад, певна величина Б, котра не веде до В. Чи обрали б ви А, котре веде до Б? Або обійдемо абстрактну філософію і зробимо так: якби вам захотілося піти до супермаркету за шоколадкою, захотілося поїхати до нього автівкою, для чого треба сісти в автівку, чи зірвали б ви двері автівки паровою лопатою? (Ні.) Інструментальна цінність (величина) — така собі «нещільна абстракція», як її називаємо ми, програмісти: іноді треба відкинути кешоване значення й обчислити дійсну очікувану корисність. Щоб бути *ефективним* і при цьому не наражатися на *самогубство*, потрібно вчасно помічати, коли зручні скорочення перестають працювати. Хоча ця формальна система і вводить інструментальні величини, вони існують лише тоді,

коли наявна необхідна постійність — виключно як зручний скорочений спосіб обчислення.

Але якщо формальну систему ускладнити до того, як ви зрозумієте просту її версію, ви можете подумати, що інструментальні величини живуть якимось своїм життям (навіть у нормативному сенсі). Тому, коли ви кажете, що Б зазвичай корисне через його зв'язок із В, ви зобов'язалися завжди пробувати Б, навіть якщо після нього не йде В. Люди припускаються такої помилки в абстрактній філософії, навіть якщо у реальному житті вони ніколи не вириватимуть двері паровою лопатою. Тут ви можете подумати, що немає способу розробити консеквенціаліста, що максимізує інклюзивну генетичну пристосованість, бо він помре з голоду, якщо йому не прописати чітку термінальну цінність «вживати їжу». Люди припускаються цієї помилки, хоча вони ніколи б не ходили колами довкола машини, намагаючись відчинити двері, через страх так туди і не потрапити, якби у них не було термінальної цінності для відчинення дверей.

Інструментальні величини живуть (у мережевій структурі) у функції умовної ймовірності, тому вона строго залежна від переконань щодо стану речей за умови фіксованої функції корисності. Якщо я вірю, що пеніцилін спричиняє пневмонію, а відсутність пеніциліну лікує пневмонію, тоді моя оцінка інструментальної цінності пеніциліну знизиться з високої до малої. Змінюєте переконання щодо стану речей — змінюєте функцію умовної ймовірності, що пов'язує дії з передбачуваними наслідками, — інструментальні величини зміняться слідом.

Деякі моральні суперечки ведуться стосовно інструментальних цінностей, а деякі — термінальних. Якщо у суперечці ваш опонент стверджує, що заборона зброї призведе до зниження рівня злочинності, а ви говорите — до підвищення, тоді ви погоджуєтеся щодо вищої інструментальної цінності (злочинність — це погано), але розходитеся щодо того, які проміжні події призведуть до яких наслідків. Але я не вважаю, що суперечка про жіноче обрізання стосується того, як найліпше досягти спільної цінності — справедливого ставлення до жінок і їхнього щастя.

У гнівних дискусіях цю важливу різницю часто *змивають у каналізацію*. Люди, які мають спільні цінності, але розходяться в описі реальності, доходять висновку, що їхні опоненти в суперечці мусять бути соціопатами. Ніби ваш запеклий ворог, прибічник контролю обігу зброї чи її легалізації, *насправді хоче вбивати людей*, що вже неправдоподібно з погляду реалістичної психології.

Боюся: мозок людини не вельми чітко типізує відмінність між термінальними моральними цінностями та інструментальними моральними цінностями. «Ми повинні заборонити зброю» і «ми повинні врятувати життя» не *відчуваються як різні* моральні переконання, як і зір не відчувається чимось відмінним від звуку. Попри всі інші способи, якими система людських цінностей ускладнює все навколо, са́ме *це розрізнення* їй вдається звести до мішанини речей з умовною цінністю.

Щоб вичленити термінальні цінності, ми маємо оглянути цю мішанину цінних речей, намагаючись зрозуміти, котрі з них отримують свою цінність звідкілясь іще. Складний проєкт, маю сказати. Якщо ви кажете, що хочете заборонити зброю, щоб знизити злочинність, знадобиться якийсь час, перш ніж ви зрозумієте, що «знизити злочинність» не термінальна цінність, а вища інструментальна, що кріпиться до цінності людського життя і щастя. І тоді прибічник права на зброю може пов'язувати це з вищою інструментальною цінністю «зниження злочинності», а також із цінністю «свобода», котра може бути для нього термінальною або ж іще однією інструментальною...

Ми не можемо видати повну систему цінностей, отриманих з інших цінностей. Імовірно, ми навіть не зберігаємо повну історію формування тих цінностей. Через розмірковування над правильними моральними дилемами вигляду «Чи зробили б ви X, якби Y?», ми часто можемо з'ясувати походження своїх цінностей. Та навіть цей проєкт ховає підводні камені — оманливі дилеми та пробіли у філософських аргументах. Ми не знаємо своїх цінностей, їхнього походження і не можемо з'ясувати, крім як братися за вразливі до помилок проєкти когнітивної археології.

Уже свідоме формулювання різниці між «термінальними цінностями» та «інструментальними цінностями», відстежування наслідків цієї різниці і коректне її використання — важка робота. І тільки при пильному погляді на просту формальну систему ми можемо побачити, наскільки легкою вона має бути у принципі.

І це я вже мовчу про всю решту ускладнень винагороджувальної системи у людей — про увесь механізм структури підкріплення: поїдання шоколаду приносить задоволення, очікування шоколаду теж приносить задоволення, хоч це й різні види задоволень...

Але я не дуже жаліюся на цей безлад.

Можливо, не відати своїх цінностей не завжди класно, але хоч життя не здасться нудним.

Нещільні узагальнення

Яблука можна їсти? Зазвичай можна, але деякі з них гнилі.

У людей десять пальців? У більшості з них так, хоч і багато хто втратив пальця, а проте вважається «людиною».

За винятком рівня опису об'єктів, що значно менші за будь-які макроскопічні: менших за суспільство, людину, палець, сухожилля, кістку і клітину, — усе нижче і нижче аж до частинок і полів, де закони дійсно універсальні, практично кожне узагальнення у реальному світі буде нещільним.

(Хоча, звісно, до вищенаведеного правила можуть бути винятки...)

Здебільшого нещільним узагальненням дають раду так: ну, доводиться просто з ними миритися. Якщо магазин солодощів майже завжди зачиняється о 22:00, за винятком Дня подяки, коли він працює до 18:00, а сьогодні ще й Національний день геноциду корінних американців, то за печивом вам краще навідатися до 18:00.

Нашій здатності працювати з нещільними узагальненнями протистоїть *потреба у завершеності* тією мірою, якою ми хочемо раз і назавжди сказати, що у людей є пальці, і дратуємося, коли доводиться миритися з постійною неоднозначністю. Внаслідок підвищення ставок може зрости потреба у завершеності, що вимикає толерантність до складності, коли вона потрібна найбільше.

Навіть якби об'єкти бажань були простішими (а вони такими не є), життя все одно було б складним. Із самої нещільної структури дійсного світу неминуче витікає ота нещільність узагальнень щодо подальших дій.

Іншими словами, інструментальні цінності часто не мають конкретних описів, котрі були б одночасно компактними і локальними.

Уявімо скриньку з мільйоном доларів усередині. Скринька замкнена не якимось звичайним кодовим замком, а десятком

клавіш, що запускають механізм відмикання. Якщо ви знаєте, як працює механізм, зможете вивести послідовність клавіш, якою відімкнете скриньку. Існує більше однієї послідовності клавіш, що активує кнопку. Але якщо ви складете достатньо хибну комбінацію, механізм спалить гроші. Якщо ж ви *не знаєте* принцип роботи механізму, ніякого простого правила штибу «Після триразового натискання будь-якої клавіші скринька відмикається» або «Після одноразового натискання п'яти різних клавіш машина спалює гроші» немає.

Існує один *компактний нелокальний* опис того, які клавіші ви хочете натиснути, щоб вони відімкнули скриньку. Можна написати компактну комп'ютерну програму, що обчислюватиме, які комбінації клавіш хороші, погані або нейтральні, але ця програма потребуватиме ще й опису машини — не тільки самих клавіш.

Так само існує і *локальний некомпактний* опис того, які клавіші натискати — велетенська довідкова таблиця результатів для кожної можливої комбінації клавіш. І ця величезна комп'ютерна програма згадує лише клавіші.

Але не існує способу описати, які комбінації клавіш хороші, погані чи нейтральні, щоб він був і *простим*, і сформульованим лише *стосовно самих клавіш*.

Тим гірше, якщо є ризик спокуситися локальними узагальненнями, котрі виявляться *нещільними*. Якщо натискати *більшість* клавіш тричі поспіль, скринька відімкнеться, але всього лиш одного натискання якоїсь особливої клавіші вистачить, щоб спалити гроші. Ви могли подумати: ви знайшли досконале узагальнення — локально описуваний клас послідовностей, після введення якого скринька відчиняється *завжди*, — коли ж ви просто не змогли уявити всі можливі шляхи функціонування механізму або оцінити всі побічні ефекти.

Цей механізм уособлює складність дійсного світу. Відкрита скринька (що є добре) і спалювання (що є погано) уособлюють тисячі уламків бажання, що складають наші термінальні

(кінцеві) цінності. Клавіші втілюють дії, стратегії, манеру поведінки, до яких ми можемо вдатися.

Якщо подумати, у скільки різних способів ми оцінюємо результати й наскільки складні шляхи їх реалізації ми обираємо, взагалі диво, що існує щось таке, як помічна етична *порада*. (Серед них найдивніша з усіх, *та все ж помічна*, — «Мета не виправдовує засобів».)

Але складність дій, навпаки, не обов'язково свідчить про складність цілей. Нерідко вам траплятимуться люди, які лукаво усміхатимуться, кажучи: «Ну, мораль — складна штука: знаєте, жіноче обрізання — це правильно в одній культурі та неправильно в іншій, зрештою, не завжди зле катувати людей. Наскільки ж ти наївний, скільки ж в тобі потреби у завершеності, що ти думаєш, що є якісь прості правила».

Ви можете сказати беззастережно і категорично, що вбивати *когось* — величезна частка негативної термінальної корисності. Так, навіть Гітлера. Утім, це не означає, що ви не маєте стріляти в Гітлера. Це означає, що чиста інструментальна корисність від розстрілу Гітлера несе величезну порцію негативної корисності від Гітлерової смерті і значно більшу порцію позитивної корисності від усіх інших життів, які було врятовано внаслідок цього.

Багато хто припускається цієї помилки типів, проти якої я застерігав ув есеї «Термінальні цінності та інструментальні цінності», і думає: якщо визнати, що чиста наслідкова очікувана корисність від Гітлерової смерті позитивна, то пряма локальна термінальна корисність також мусить бути позитивною — тобто моральний принцип «Смерть — це завжди погано» сам по собі є нещільним узагальненням. Але це подвійний підрахунок з корисностями замість імовірностей — ви створюєте резонанс між очікуваною корисністю і корисністю замість одностороннього переходу від корисності до очікуваної корисності.

Чи, може, це таке прагнення до односторонніх політичних дебатів: найкраща стратегія не повинна мати *жодних* недоліків.

У моїй моральній картині світу *локальна* корисність Гітлерової смерті стабільно негативна, хай там якими будуть зовнішні

наслідки і, відповідно, *очікувана* корисність.

Ясна річ, ви можете навести моральний аргумент, що *само собою* добре карати злих людей, навіть якщо це смертна кара для досить злих людей. Але вам не вдасться обґрунтувати цей аргумент, вказуючи на те, що *наслідком* пострілу у чоловіка з націленою зброєю може бути порятунок інших життів. Цей аргумент посилається на цінність життя, а не смерті. Якщо очікувані корисності нещільні і складні, самі корисності не обов'язково такі ж нещільні і складні. Можуть бути такими! Але це вже тема для окремої розмови.

Прихована складність бажань

(Мені стало відомо, що зараз цю статтю помилково інтерпретують як доказ того, що я (чи Інститут машинного інтелекту) раніше обстоював, що доносити до машинних суперінтелектів людські цінності чи навчати їх передбачати ті цінності буде вкрай складно. Очевидно, це не так, і до того ж я писав не про те. Приклад, описаний ув есеї, буквально взагалі не стосується ніякого штучного інтелекту! Якби у дописі йшлося про те, чого ШІ нібито не може робити, у центральному прикладі я б написав про ШІ! Теза, представлена нижче, стосується алгоритмічної складності людських цінностей. Ця теза актуальна у рамках ширшого аргументу, оскільки вона стосується складності того, що потрібно, аби штучний суперінтелект чогось ***прагнув*** *або* ***цінував****, а не того, що суперінтелект нібито не зміг би* ***передбачити*** *чи* ***зрозуміти****. — ЕЮ, травень 2024)*

<

> «Я хочу жити там, де сам вирішив, у фізично здоровій непошкодженій і зовні звичайній версії мого поточного тіла з моїм нинішнім ментальним станом; я хочу тіло, що загоюватиметься після усіх поранень на три сигми швидше, ніж у середньому, з огляду на доступні мені медичні технології, що буде захищене від усіх хвороб, травм чи недуг, що ведуть до інвалідності, болю чи погіршеного функціонування будь-якого відчуття, органу чи функції тіла понад десять днів поспіль або п'ятнадцяти днів протягом будь-якого року...» -- Проєкт Open-Source Wish[1], Бажання безсмертя 1.1[2]

1. https://web.archive.org/web/20120627051804/http://homeonthestrange.com/phpBB2/viewforum.php?f=4

2. https://web.archive.org/web/20111217210507/http://homeonthestrange.com/phpBB2/viewtopic.php?t=95

Існують три види джинів. В одних ви можете сміливо просити: «Я хочу, щоб ти зробив те, що я маю попросити у тебе». Другим — озвучувати бажання *не*безпечно, хай чого ви бажаєте. Треті не настільки сильні чи розумні.

Скажімо, ваша старенька мати застрягла у палаючому будинку, а ви, за збігом обставин, прикуті до візка, тому ніяк не зможете кинутися їй на допомогу. Ви уже й кричали: «Дістаньте маму з пожежі!», але ніхто не почув.

На щастя, у вашій кишені якраз завалявся генератор результатів. Цей зручний пристрій стискає потік часу, наповнюючи ймовірністю одні події, висмоктуючи її з інших.

Генератор результатів не має свого розуму. Він обладнаний крихітною машиною часу, що повертає час назад, *якщо* вказаний результат *не* настає. Наприклад, якщо ви під'єднали сенсори генератора результатів до монетки і налаштували його так, щоб машина повертала час, поки монетка не випаде гербом, а потім *ви* власне підкидаєте монетку і бачите, що вона дійсно випадає гербом. (Фізики скажуть, що будь-яке майбутнє, в якому відбувається «перезапуск», суперечливе, а отже, такого майбутнього взагалі ніколи не відбувається, тому насправді ви не вбиваєте ніяку версію себе.)

Хай яке твердження ви вводите у генератор результатів, *якимось чином* воно реалізується — за умови, що воно не порушує законів фізики. Якщо спробуєте ввести твердження, яке *надто* малоймовірне, машина часу тут же — ще до настання результату — дасть механічний збій.

Крім того, ви можете і перенаправити потік імовірностей більш кількісними методами, використовуючи «функцію майбутнього» для масштабування ймовірності часового перезапуску для різних результатів. Якщо ймовірність часового перезапуску складає 99%, коли монетка випадає гербом, і 1%, коли — числом, шанс перерозподілиться з 1:1 до 99:1 на користь числа. Якби у вас була загадкова машина, що видає гроші, і ви хотіли максимізувати число виданих грошей, ви б застосовували ймовірності перезапуску, що зменшуються зі зростанням суми

грошей. Наприклад, видача десяти доларів може складати ймовірність перезапуску 99,999999%, а видача ста доларів — 99,99999%. Так ви зможете отримати результат, що, як правило, буде якомога вищим у функції майбутнього, навіть якщо найкращий досягнутий максимум вам невідомий.

І от ви відчайдушно висмикуєте з кишені генератор результатів — пам'ятаєте ж, що ваша мати досі у пастці палаючого будинку? — і пробуєте сформулювати свою ціль: *витягнути вашу маму з того будинку*!

Користувацький інтерфейс не приймає запитів англійською[3] мовою. Пам'ятаєте, що генератор результатів не має власного розуму? Але у ньому вбудовані 3D-сканери ближньої дії і утиліти для зіставлення зі зразком. Ви тримаєте світлину вашої матері з її головою і плечима, показуєте генератору для зіставлення, застосовуєте метод суміжності об'єктів, щоб виділити все тіло матері (а не лише її голову та плечі) і за допомогою відстані від центру будинку до вашої матері задаєте *функцію майбутнього*. Що далі від центру, то менша ймовірність перезапуску машини часу.

Про всяк випадок ще раз кричите: «Витягни мою маму з того будинку!» і натискаєте «Ввести».

На якийсь момент здається, що нічого не відбувається. Роззираєтеся. Ви чекаєте, коли під'їде пожежна автівка, прибудуть рятівники або звідкілясь навіть прибіжить бігун-атлет, який дістане вашу матір з будинку.

БУУМ! Під будинком оглушливим ревом вибухає головна газова магістраль. Поки уся ця палаюча конструкція розпадається, ніби в режимі сповільненого перегляду ви бачите, як понівечене тіло вашої матері, пришвидшуючись, злітає у повітря, все більше віддаляючись від того, що колись було центром.

У генератора збоку є кнопка екстреного шкодування. У всіх функціях майбутнього натисканню кнопки шкодування автоматично приписується величезне від'ємне значення — імовірність

3. Або будь-якою іншою природною мовою. — Прим. пер.

часового перезапуску становить приблизно 1, — тож генератор результатів майже напевно не зробить нічого такого, що настільки засмутило б користувача, аби той натиснув кнопку шкодування. Ви просто не матимете можливості згадати, що колинебудь натискали її. Але заледве ваш палець встиг торкнутися тієї кнопки (яка ж тепер з неї користь?), з неба на вас падає і розчавлює розжарена дерев'яна балка.

Ви точно хотіли не цього, але у визначеній функції майбутнього такий результат набирає багато балів...

Генератор результатів — джин другого типу. Жодне бажання *не* є безпечним.

Якби вас хтось попросив дістати їхню стареньку матір з палаючого будинку, ви могли б допомогти або вдати, що нічого не чуєте. Але вам і *в голову не прийшло б* підірвати будинок. Прохання «Дістань мою маму з того будинку» *звучить* набагато безпечніше, ніж воно насправді є, оскільки ви навіть не *розглядаєте* тих планів, яким призначаєте екстремальні від'ємні значення.

Подивімося ще раз на трагедію групового добору. Дехто з ранніх біологів стверджував, що груповий добір для малих розмірів субпопуляцій призведе до індивідуальної стриманості у розмноженні. І все ж практичне застосування групового добору у лабораторних умовах призводило до канібалізму, особливо до поїдання статево незрілих самиць. У ретроспективі нам уже очевидно, що за умов жорсткого добору у групах малих субпопуляцій канібали матимуть більші репродуктивні успіхи, ніж особини, що добровільно відмовилися від репродуктивних можливостей. Але поїдання статево недозрілих самиць виявилося настільки *неестетичним* рішенням, що Вінн-Едвардсу, Аллі, Бреретону та іншим прихильникам групового добору воно просто не спало на думку. Вони бачили лише ті рішення, якими скористалися б самі.

Нехай ви пробуєте залатати функцію майбутнього своїм уточненням, що генератор результатів не повинен підривати будинок: множина результатів, у яких будівельні матеріали надто

розпорошені в просторі, матиме ймовірність часового перезапуску ~1.

І в цій гілці ваша мати випадає з вікна і ламає шию. Генератор результатів узяв інший курс крізь час, що все ще обернувся розташуванням матері поза будинком, але не так, як ви хотіли, а людині-рятівникові такий розв'язок і не спав би на думку.

Якби ж то проєкт Open-Source Wish[4] розробив бажання, котре дістало б вашу матір з будинку у вогні:

> «Я хочу перенести мою маму (жінку, яка передала мені половину своїх генів і народила мене) за кордони палаючого будинку, що наразі найближче до мене; але не треба підривати той будинок, не треба розносити стіни аж до стирання кордонів того будинку, не треба чекати, поки приїдуть рятівники і дістануть її тіло уже після того, як будинок згорить до фундаменту...».

Усі ці особливі випадки — здавалося б, нескінченна кількість потрібних латок — мають нагадати вам притчу про штучне додавання — програмування систем фахівців з арифметики через ручне додавання все більшої кількості тверджень типу «п'ятнадцять плюс п'ятнадцять дорівнює тридцять, але п'ятнадцять плюс шістнадцять дорівнює тридцять один».

Як *вам* виключити результат, за якого будинок вибухає і підкидає вашу мати у небо? Ви дивитеся наперед і розумієте, що вона все одно помре; такого наслідку ви не хочете, тож намагаєтеся заборонити подію, що призвела до цього.

Ваш мозок не запрограмований на конкретне, заздалегідь записане твердження «Підривати будинок у вогні, всередині якого моя мати, — погана ідея». І все ж ви намагаєтеся записати са́ме це конкретне твердження у функції майбутнього генератора результату. Внаслідок цього складність бажання стрімко зростає, перетворюючись на велетенську таблицю пошуку, що

4. https://homeonthestrange.com/phpBB2/viewforum.php?f=4

записує ваше судження стосовно кожного можливого шляху у часі.

Ви не попросили того, чого дійсно хочете. *Хотіли*, щоб ваша мати лишилася живою, а *побажали*, щоб вона якомога віддалилася від центру.

Хіба що і це не все, чого ви хотіли. Якби вашу мати врятували з будинку, але на її тілі лишилися тяжкі опіки, у вашому рейтингу вподобань такий результат стояв би нижче за порятунок її цілою і неушкодженою. Отже, ви цінуєте не лише життя своєї матері, а й здоров'я.

І не лише фізичне здоров'я, а й ментальне. Буде гірше, якщо її врятують способом, який їй нашкодить (наприклад, велетенське пурпурове чудовисько з'являється нізвідки і забирає її), на відміну від того, якби з'явився пожежник та евакуював її туди, куди ще не дістався вогонь. (Так, нам слід дотримуватися законів фізики, але, може, досить потужний генератор результатів здатен дістати іншопланетян, які випадково гуляли околицями у той самий момент.) Однак, ви, певна річ, надасте перевагу порятунку монстром, а не запіканню живцем.

А якщо раптово відкриється кротовина і викине її на безлюдний острів? Краще за перспективу її смерті; гірше за перспективу бути живою, здоровою, неушкодженою і в постійному контакті з вами та іншими членами її кола соціальних зв'язків.

Чи погодилися б ви, якби вашу мати врятували ціною життя вашого домашнього собаки, коли той, гавкаючи, біг до пожежника і його переїхала автівка? Ясно, що так, але за інших рівних умов було б краще, якби собака лишився живим. Обміняти життя іншої людини на неї вам би не захотілося, але як щодо життя засудженого вбивці? Чи щось зміниться, якщо вбивця помре при спробі зі своєї доброти врятувати вашу матір? А двоє вбивць? А якщо ціною порятунку матері буде знищення кожної збереженої копії «Маленької фуги соль мінор» Баха — і навіть спогадів про неї? А якби вона була невиліковно хворою і зрештою померла за вісімнадцять місяців?

Якщо її ногу розчавила розжарена балка, чи варто діставати решту її тіла? А якщо їй розчавило голову? А якщо їй розчавило все тіло, крім голови? Що як поряд уже чекає команда кріоністів, готова забрати її голову? Чи заморожена голова рахується за людину? А Террі Шайво — людина? А наскільки цінне шимпанзе?

Ваш мозок не нескінченно складний — існує лише скінченна колмогорова складність / довжина повідомлення, якої вистачить для опису усіх зроблених вами суджень. Однак її скінченність не робить її короткою. Ми цінуємо багато чого, що не зводиться до щастя чи репродуктивної пристосованості.

Не існує безпечного бажання, меншого за сукупність людської моралі. Забагато можливих шляхів крізь Час. Ви не зможете уявити всі можливі дороги, що ведуть до місця призначення, котре ви задали джину. Ще ефективніше «максимізувати відстань між вашою матір'ю та центром будинку» можна підривом ядерної боєголовки. Або джин усією своєю силою може викинути її тіло за межі Сонячної системи. Або, якщо джин виявиться надрозумним, він зробить щось таке, про що ні ви, ні я і не подумали б — як шимпанзе не подумало б про підрив ядерної боєголовки. Ви не можете уявити всі ходи в часі, як не можете запрограмувати шахову машину, якій задаєте хід для кожної можливої позиції на дошці.

А життя куди складніше за шахи. Заздалегідь ви ніяк не передбачите, котрі цінності знадобляться, щоб судити про шлях, яким піде джин. Тим паче, якщо ви бажаєте чогось більш довгострокового чи широкомасштабного за порятунок вашої матері з будинку у вогні.

Боюся, проєкт Open-Source Wish марний, хіба що ми тлумачимо його як ілюстрацію того, як *не* потрібно осмислювати джинові задачі. Єдиний безпечний джин — той, що поділяє ваші критерії судження, якому ви зможете просто сказати: «Я хочу, щоб ти зробив те, що я маю попросити у тебе». І це запускає функцію джина «повинен».

Насправді не потрібно навіть *чогось* говорити. Щоб джин безпечно здійснив бажання, він має поділяти ваші цінності, що привели вас до цього бажання. Інакше джин може не обрати шлях у часі, що веде до задуманого вами місця призначення, або не виключити жахливих побічних ефектів, через які ви й поготів не розглядатимете план. Бажання являють собою нещільні узагальнення, виведені з величезної, але скінченної структури вашої сукупної моралі, тож заткнути ці щілини ви можете лише, включивши усю цю структуру.

Якщо ваш джин безпечний, зайве навіть бажати. Просто запустіть джина.

Антропоморфний оптимізм

Основна помилка антропоморфізму — очікувати, що щось нелюдське можна передбачити чорною скринькою нашого мозку, тоді як причиново-наслідкова структура нелюдського настільки відрізняється від структури людського мозку, що не дає вам жодних підстав цього очікувати.

Трагедія групового добору (про яку я вже писав у ланцюжку про еволюцію) була, скоріше, надзвичайною помилкою, допущеною групою ранніх (до 1966) біологів, включно з Вінн-Едвардсом, Аллі та Бреретоном, котрі вірили, що хижаки добровільно обмежуватимуться у розмноженні, щоб уникнути перенаселення їхнього середовища проживання і вичерпування популяції здобичі.

Запропонована теорія звучала так: якщо існують численні географічно віддалені *групи*, скажімо, лисиць, ті групи, які найкраще контролювали свою репродуктивну активність, відправляли б колоністів для заміщення вимерлих популяцій. І так з часом ідея про груповий добір поширить гени утримання від розмноження серед лисиць.

Я не повторюватиму всіх тих проблем, що склалися у цьому сценарії. Досить і того, що початково не було жодних емпіричних свідчень; таких свідчень так і не знайшли; чисельність популяцій хижаків постійно різко скорочується; і до того ж математично майже неможливо, щоб тиск групового добору переважав супротивний тиск індивідуального.

Оскільки теорія виявилася цілковито хибною, ми можемо запитати, чи батьки теорії робили щось не так.

«Навіщо бути таким грубим? — запитаєте ви. — Як вони могли знати до початку експерименту, що груповий добір не зможе побороти індивідуальний?».

Пізніше Майкл Вейд *провів-таки* експеримент і створив у лабораторії умови для групового добору, близькі до неможливих. Для низької чисельності Вейд багато разів добирав

субпопуляції комах. Чи еволюціонували комахи до самообмеження у розмноженні, жили у мирі й достатку їжі, як і передбачали прихильники теорії групового добору?

Ні. Дорослі особини пристосувалися пожирати власні яйця й личинки (особливо личинки жіночих особин).

Певна річ, у процесі добору особини у невеликих субпопуляціях не обмежують *себе* у розмноженні. Такий добір призводить до появи особин, які їдять дітей *інших*. Особливо дітей-самиць.

Чому ж прихильникам групового добору *не* спала на думку така можливість?

Уявімо, що ви живете у племені. Ви дізнаєтеся, що у найближчому майбутньому вашому племені перестане вистачати ресурсів. Тож ви, мабуть, запропонуєте кожному подружжю заводити по одній дитині: після першої подружжя переходить на презервативи. Вам і *на думку не спаде* сказати щось штибу «А давайте кожен з нас розплодить якомога більше виводку, а потім ми будемо полювати на дітей одне одного і поїдати їх, особливо дівчаток».

Подумайте про порядок пріоритетності рішень відповідно до ваших цілей. Ви хочете, щоб рішення стояло у рейтингу вподобань якомога вище. Як знайти таке рішення? Мозком, звісно! Подумайте про свій мозок як про *генератор високопріоритетних рішень* — процес пошуку, котрий породжує рішення, які займають вищі позиції у вашому внутрішньому рейтингу вподобань.

Простір рішень для усіх реальних проблем загалом вельми широкий — для цього і потрібен *ефективний* мозок, який навіть не *обтяжує себе окресленням* абсолютної більшості низькопріоритетних рішень.

Якщо ваше плем'я стикнулося з нестачею ресурсів, ви можете спробувати стрибати всюди на одній нозі або повідкушувати собі пальці на ногах. Такі «рішення», очевидно, не робочі й понесуть за собою значні втрати, та й ефективність мозку просто не дозволить марнувати час на обдумування таких кепських рішень — він і не видасть їх. У пошуку високопріоритетних

рішень ваш мозок відразу ж звертається до частин простору рішень на кшталт «Усі члени племені гуртуються і погоджуються заводити не більше однієї дитини на подружжя, поки ресурси не відновляться».

Таке *низькопріоритетне* рішення, як «Кожен з нас розплодить якомога більше виводку, а потім ми будемо поїдати дівчаток», навіть не *виникне у процесі пошуку*.

Однак позначення одного з варіантів як «низького» чи «високого» не є властивістю самого варіанта, а процесу оптимізації, що визначає пріоритетність. Тож різні оптимізаційні процеси виконуватимуть пошук у різних порядках.

У контексті *еволюції* немає нічого дивного в особинах, котрі розмножуються по максимуму, а потім поїдають чиїхось доньок, тоді як стратегія добровільного самообмеження заради блага групи геть абсурдна. У популяції перший набір алелів швидко замінить другий, якщо говорити менш антропоморфно. (І тут в природного добору немає очевидного порядку пошуку — ці дві альтернативи здаються настільки ж простими, як і мутації.)

Скажімо, хтось із біологів колись сказав: «Якщо у популяції хижаків лише обмежені ресурси, еволюція пристосує їх до добровільного самообмеження у розмноженні — саме так зробив би *я*, якби *я* був відповідальним за створення хижаків». Ось це — антропоморфізм у чистому вигляді. Ланцюжки міркувань видно відразу: *я* б зробив отак, тому я роблю висновок, що й *еволюція* зробить це отак.

У моїй роботі доводиться стикатися з цією хибою у її чистому вигляді. Але хай навіть ви скажете: «ШІ не обов'язково працюватиме так, як ви». Нехай ви скажете і цьому гіпотетичному біологові: «Еволюція працює не так, як ви». Що він скаже у відповідь? Скажу вам, що відповідь *не* звучатиме так: «Боже правий! Я ж не розумів цього! Котрийсь із етапів мого міркування був помилковим; я викину висновок і почну спочатку».

Ні. *Натомість* ви почуєте, що будь-яка система ШІ має міркувати так само, як і мовець, або що природний добір, виходячи

з геть інших оптимізаційних критеріїв і застосовуючи геть інакші оптимізаційні методи, має робити *саме те*, що людині подумалося б як гарна ідея.

Звідси й витончена ідея, що груповий добір буде прихильний до тих груп хижаків, які добровільно відмовилися від можливості розмножуватися.

У своїх помилкових прогнозах прихильники групового добору дали маху не менше від тих, хто безпосередньо вдається до цієї помилки. Їхні кінцеві висновки нічим не відрізнялися від висновків, до яких вони б дійшли, якби припустили, що еволюція думає так, як вони. Вони стерли те, що було написано *зверху* нижнього рядка їхнього аргументу, але *не* стерли власне нижній рядок — зате записали нові раціоналізації. Ось так і маскується хибне міркування; *очевидно* хибний крок у процесі висновування приховали — хоча висновок не змінився; отже, на ділі він точно так само хибний.

Але навіщо науковцю робити таке? Зрештою ж, дані суперечили погляду прихильників групового добору — вони збентежилися.

Як я зазначив в есеї «Критерії помилкової оптимізації», люди, видно, розвинули інстинкт стверджувати, що обрана *нами* стратегія випливає практично з *будь-якого* оптимізаційного критерію. Політика була рисою середовища предків. Ми — нащадки тих, хто найпереконливіше обґрунтовував, що в інтересах *племені* — не просто у їхніх особистих — стратити ненависного племінного суперника Уґлака. Ми явно не нащадки Уґлака, який не зумів довести: моральний кодекс племені вимагав залишити його в живих, а не лише через його очевидний особистий інтерес.

І з огляду на наше вміння наводити переконливіші аргументи (за те, у що щиро віримо), ми розвинули інстинкт щиро вірити, що цілі інших людей і моральний кодекс нашого племені насправді передбачають, що заради *їхнього* ж блага вони мають робити все по-*нашому*.

Так і прихильники групового добору, науявлявши красиву картинку того, як хижаки обмежують себе у розмноженні, інстинктивно раціоналізували, чому природний добір має працювати *по-їхньому*, навіть відповідно до цілей природного добору. Ті лисиці будуть більш пристосованими, якщо обмежуватимуть себе у розмноженні! Ні, справді! Вони навіть розплодяться більше за інших лисиць, які не обмежують себе! Чесно!

Тільки-от спроби переконати еволюцію працювати по-нашому програшні тим, що в еволюції немає позиції, яку ви змогли б зсунути своїми аргументами. Еволюція не працює так, як ви, — всередині неї немає навіть якоїсь частини, яка б прислухалася чи *зважала на* ваші розлогі пояснення, чому вона має робити усе по-вашому. Аргументи людей навіть не *вкладаються* у внутрішню структуру природного добору як оптимізаційного процесу: не людськими аргументами еволюція сприяє поширенню алелів, на відміну від політики, де наші аргументи відіграють причиново-наслідкову роль.

Тож, не *спромігшись* переконати природний добір працювати так, як вони сказали, прихильники групового добору зніяковіли, коли реальність виявилася іншою.

Тут присутній досить глибокий підтекст стосовно Недружнього ШІ.

Якщо узагальнити, то у цьому й полягає проблема оптимістичного міркування загалом. Що таке оптимізм? Ранжування можливостей за нашим рейтингом уподобань та обрання результату, що стоїть вище у тому рейтингу і потім якимось чином стає вашим передбаченням. Які саме вишукані раціоналізації створювалися на шляху, певно, немає значення — достатньо подивитися на історію людського пізнання: оптимізм на вході — оптимізм на виході. Але Природа чи будь-який інший процес, про який іде мова, *наспрaвді* не обирає у *причиновому* сенсі між варіантами, ранжуючи їх за вашим рейтингом вподобань й обираючи один із вищих у цьому порядку. Отже, саме мозок не може узгодитися із навколишнім середовищем, і тому передбачення не збігається з реальністю.

Втрачені цілі

Як я ще ходив чи то у дитсадок, чи то у перший клас, мені дали транслітеровану єврейську молитву і вперше просили помолитися. Я запитав значення слів тієї молитви. Мені відповіли, що поки я молюся івритом, знати значення слів було не обов'язково — молитва спрацює так чи інакше.

Так я розійшовся з юдаїзмом.

Поки ви читаєте есей, якийсь хлопець чи дівчина сидить за університетською партою і старанно вивчає матеріал, який не збирається застосовувати і до якого не має внутрішнього інтересу. Студенти хочуть високооплачувану роботу, яка, своєю чергою, вимагає шматка паперу, для якого потрібен попередній ступінь магістра, а для цього — ступінь бакалавра, а університет, що дає той ступінь, потребує від вас проходження курсів техніки в'язання XII століття. Тож вони ретельно навчаються, щоб після кінцевого іспиту зрештою забути вивчене, але і на тому не зупиняються, бо вони ж так *хочуть* того папірця.

Може, *ви* й розуміли, яким безглуздям займалися, але навчання не кидали. У вас просто не було вибору. Так ви собі казали?

Згідно з недавнім дослідженням[1], проведеним тут, в агломерації затоки Сан-Франциско, 80% вчителів початкових класів (K-5) приділяють природничим наукам менше години на тиждень, 16% взагалі не виділяють на них час. Як таке можливо? Як я розумію, безпосередня причина — Закон «Про недопущення відставання дітей у навчанні» (No Child Left Behind) та аналогічні законодавчі ініціативи. Тепер практично увесь класний час виділяється на підготовку до тестів, обов'язкових до складання на державному чи федеральному рівні. Пам'ять підказує (хоча джерела згадати не можу), що в одній зі шкіл лише на *складання* обов'язкових тестів ішло 40% навчального часу.

1. https://www.sfgate.com/bayarea/article/Science-courses-nearly-extinct-in-elementary-3236187.php

Стара радянська бюрократія славилася більшою любов'ю до показовості, ніж до реальності. Одна взуттєва фабрика[2] перевиконала план, виготовивши велику партію взуття малих розмірів. Інша взуттєва фабрика прозвітувала про виготовлення «взуття», хоча то були порізані, але незшиті шматки шкіри. Вищі чини бюрократії не дуже й хотіли надто ретельно розслідувати ситуацію — вони хотіли звітувати про перевиконання плану. Яка неоціненна допомога для товаришів, які відморожували собі ноги!

У низці джерел висловлюється думка, що більшість опублікованих досліджень[3] у медицині, попри статистичну значущість ($p < 0,05$), насправді неправдива. Але поки поріг для публікації лишається на рівні $p < 0,05$, навіщо комусь дотримуватися вищих стандартів, якщо для цього потрібні великі гранти — щоб збільшувати експериментальних груп, — а це знижує ймовірність публікації? Усі ж і так знають, що *уся наука* зводиться до публікації безлічі статей, *університет* — до випуску шматків пергаменту, а *школа* — до складання обов'язкових іспитів, які гарантують річний бюджет. Встановлювати правила гри не випадає, а грати за іншими правилами — означає просто програти.

(Хоча з якихось причин наукові видання з фізики вже вимагають порогу $p<0,0001$. Видається, у тих редакціях придумали якусь іншу мету, крім власне публікації.)

У супермаркет завезли шоколад. До нього можна доїхати автівкою. Доїхати вдасться, якщо сидіти у машині. А машину треба відчинити ключем. Тож вам потрібен ключ. Якщо виявиться, що шоколад не доїхав до супермаркету, ви не будете стояти біля автівки, клацаючи зовнішньою ручкою (бо ж треба відчинити двері!). Нечасто я бачив, щоб люди губилися у планах, які придумували самі.

Інше діло — коли мотивація мусить проходити через великі організації, ба навіть гірше — багато різних організацій та зацікавлених груп, причому деякі з них державні. Тоді ви

2. https://wiki.c2.com/?SovietShoeFactoryPrinciple

3. https://www.researchgate.net/publication/7686290_Why_Most_Published_Research_Findings_Are_False

спостерігаєте прояви поведінки, які, крім як «повне божевілля», і не описати, якби їх породив єдиний розум. Комусь платять лише за те, що вони відкривають двері автомобіля, тому що цей показник можна виміряти, і цю людину не хвилює, чи водієві заплатять за те, що він доїхав до супермаркету, не кажучи вже про те, чи клієнт придбає шоколад, чи його споживач лишиться задоволеним.

З погляду баєсіанства підцілі являють собою побічні явища функцій умовної ймовірності. Без корисності годі думати про очікувану корисність. Чи не було б безглуздо думати, що інструментальна цінність може жити своїм *математичним* життям, залишивши далеко позаду термінальну? За мірками теорії рішень це цілковите безглуздя.

Але перейдімо до Закону «Про недопущення відставання дітей у навчанні». Політики хочуть створювати враження, ніби вони розв'язують освітні проблеми; *цього* року політики мають створювати видимість зайнятості в очах виборців цього року, а не через п'ятнадцять років, коли ті діти шукатимуть роботу. Зрештою, політики не є кінцевими споживачами освіти. Бюрократи змушені демонструвати зріст — вони зацікавлені у прогресі, який вимірюється цим роком. Це не вони залишаться без знань у природничих науках. Не видавці, які випускають підручники, і не комісії, що їх закуповують, нудитимуться у класах.

Кінцеві споживачі знань — діти, які не можуть заплатити, проголосувати і засідати в комітетах. Хоч їхні батьки й піклуються про них, вони не можуть постійно сидіти з ними на уроках — лише притягати політиків до відповідальності на підставі поверхневих уявлень про «жорсткий підхід до освіти». Політики надто зайняті переобранням, щоб самим вивчати усі дані, тож їм лишається спиратися на поверхневі звіти від бюрократів, які зайняті замовленням досліджень, — усе це не обов'язково допомагає дітям, зате дозволяє політикам зберігати образ піклувальників. Бюрократи знають, що не користуватимуться підручниками, тому й не задумуються, наскільки зручно їх читати дітям — аби процес закупівлі відповідав регламентованій

процедурі. У видавців немає мотиву випускати *неякісні* підручники, але вони знають, що комітет із закупівлі буде порівнювати їх за кількістю охоплених тем, що комітет із закупівлі для четвертого класу не координується із комітетом по третьому класу, тому вони втискують якомога більше тем в один підручник. До кінця року вчитель не встигає пройти і чверті підручника, а наступного року починається те саме. Хоч учителі й можуть скаржитися, ухвалюють рішення не вони, та й зрештою не їхнє майбутнє стоїть на кону, що різко обмежує кількість вкладених зусиль за неоплачуваний альтруїзм...

Якщо подивитися на все це під таким кутом (подумайте, скільки втрачається інформації і мотивації), дивовижно, що від початкової мети — здобути знання — лишається бодай щось. А втім, багато освітніх систем, здається, переживають занепад і перетворюються на щось таке, що не краще за повну відсутність освіти.

Хочете, щоб проблемі знайшли *справжнє* рішення? Відправте тих політиків до школи.

Один людський мозок здатен відстежувати ймовірнісне очікування корисності, яке проходить через ланцюжок умовних ймовірностей десятка проміжних подій, включно з нелокальними залежностями, коли, наприклад, очікувана корисність відкриття дверей залежить від наявності шоколаду в супермаркеті. Однак організації сьогодні можуть винагороджувати лише за те, що сьогодні можна виміряти і вписати у юридичний контракт, — тобто замість далекосяжних наслідків вони вимірюють проміжні події. А проміжні показники, своєю чергою, є нещільними узагальненнями — і часто дуже нещільними. Бюрократи — це такі собі джини, яким не можна довіряти, позаяк вони не розділяють цінностей того, хто загадує бажання.

Якось Міямо́то Муса́ші сказав таке:

> «Коли ви берете до рук меча, головне — ваш намір розрубати противника будь-якими засобами. Щоразу, коли ви блокуєте удар чи атакуєте, наскочуєте, рубаєте чи

> торкаєтеся меча супротивника, тим самим рухом ви повинні розрубати його. Саме розрубати його шкіру. Бо якщо ви думаєте лише про те, як вдарити, наскóчити, атакувати чи торкнутися його меча, ви не зможете пройтися мечем по його тілу. Понад усе ви маєте думати про те, як своїми рухами прорубатися до противника, розрубавши його. Це питання варто ретельно вивчити».[4]

(Хотів би *я* жити у часи, коли, сказавши читачам ретельно дослідити якесь питання, я їх не ображу.)

Чому взагалі під час фехтування хтось втрачає з виду ціль? Якби їх навчив битися хтось інший, якби вони не створили самі все це мистецтво з нуля, вони, мабуть, не розуміли б, навіщо в якийсь момент відбивати, в інший — наскочувати, не зуміли б збагнути, коли з правил є винятки, не помітили б, коли звичайний прийом не спрацьовує.

Розуміти, як кожнісіньке правило своїм застосуванням приводить нас до істини, — наріжний камінь епістемічної раціональності. Бачити, як кожнісінька очікувана корисність наближає нас до корисності, — відповідний наріжний камінь прагматичної раціональності (теорія рішень на противагу теорії ймовірностей). Це питання варто ретельно вивчити.

Кéролін Чéррі якось сказала таке:

> «У вашого меча немає леза. Лише ваш намір. Коли ви втрачаєте друге, ви лишаєтеся без зброї».

Я бачив, як багато хто збивався на манівці, коли звертався з проханнями до уявного джина штучного інтелекту, вигадуючи одне бажання за іншим, яке здається їм правильним, — іноді з певними застереженнями, а іноді навіть без найменшого натяку на обережність. І при цьому вони не переходять на метарівень. У них немає інстинкту дивитися на мету — інстинкту, що скерував мене, п'ятирічного, на шлях атеїзму. Немає рефлексу питати: «Чому, як на мене, це бажання вдале? А джин судитиме так

4. A Book of Five Rings: The Classic Guide to Strategy by Miyamoto Musashi

само?». Вони не бачать, звідки витікає їхнє судження, що його спричиняє. Вони втрачають з поля зору м'яч: знають, що м'яч відскочив, але не мають інстинкту подивитися від чого — що було критерієм формування судження.

Так само люди машинально не помічають, коли гадані егоїсти наводять альтруїстичні аргументи на користь егоїзму чи гадані альтруїсти — егоїстичні аргументи на користь альтруїзму.

Люди чудово відстежують цілі, наприклад, поїздки до супермаркету, коли *все* це вміщується у їхній голові і не потрібно ні джинів, ні філософій. Біда в тому, що цивілізація насправді куди складніша за поїздку до супермаркету. Між дитиною, яка страждає в класі, та новоспеченим випускником, який не дуже добре виконує свою роботу, стоять десятки організацій і десятки років. (Та чи помітить це рекрутер або менеджер, якщо випускник вміє імітувати зайнятість?) І так з кожною новою проміжною ланкою між дією та її наслідком у намірі виникає ще одна можливість відхилитися від курсу. З кожною проміжною ланкою втрачається що інформація, що мотивація. І вся ця ситуація турбує більшість людей куди менше, ніж мене; інакше чому всі мої однокласники були готові молитися, не розуміючи, що саме вони промовляють? У них не було того самого інстинкту — подивитися на причиновий механізм.

Чи можуть люди взагалі навчитися тримати погляд на м'ячеві? Не давати наміру збитися на манівці. Не наскочувати, не завдавати удару, не торкатися меча без розуміння вищої мети, до якої вони наближаються тими рухами. За інших рівних умов люди зазвичай *хочуть* зробити свою роботу. Чи може взагалі існувати розсудлива корпорація? А розсудлива цивілізація? Це лише далека мрія, але саме до неї я йшов усіма цими дописами про потік намірів (відомих як очікувана корисність чи інструментальна цінність) із цілями (також відомими як корисність чи термінальна цінність). Чи можуть люди навчитися відчувати потік від батьківських до дочірніх цілей? А розуміти — свідомо чи опосередковано — різницю між очікуваною корисністю і корисністю?

Чи переймаєтеся ви загрозами для вашої цивілізації? Найгірша метазагроза для складної цивілізації — її власна складність, оскільки це ускладнення веде до втрати багатьох цілей.

Оглядаючись назад, я бачу, що моїм життям перш за все керувала винятково сильна огида до втрачених цілей. Сподіваюся, її можна перетворити на навичку, що піддається вивченню.[5]

5. Концепція «втраченої цілі» описує ситуацію, коли люди поступово забувають, для чого насправді потрібна певна дія або правило. Спочатку якась процедура є лише засобом для досягнення важливішої мети. Але в складній системі між цією метою і людиною, яка виконує процедуру, може з'явитися багато проміжних ланок. Через це люди дедалі гірше бачать, як їхня робота пов'язана з кінцевим результатом, а кожна ланка може мати власні правила та стимули. Цей процес називають *заміщенням цілей*. Близьке до нього поняття — *закон Гудгарта*: показник, який спочатку використовували лише для вимірювання бажаного результату, перестає бути хорошим показником, коли його роблять безпосередньою метою. — Прим. пер.

Путівник словами для людей

Притча про кинджал

Був собі придворний блазень, і любив він бавитися логікою.

От якось приніс блазень королю дві скриньки. На першій написано було:

> «Або у цій скриньці сидить сердита жаба, або у скриньці із хибним написом сидить сердита жаба, але не в обидвох».

На другій написано було:

> «Або у цій скриньці золото і в скриньці з хибним написом сидить сердита жаба, або сердита жаба сидить у цій скриньці, а скринька з істинним написом містить золото».

І сказав блазень королю: «В одній скрині сидить сердита жаба, в іншій лежить золото; одна і тільки одна з них позначена істинним написом».

Король відкрив не ту коробку, і його покусала сердита жаба.

«Бачте, — почав блазень, — припустімо, що перший напис істинний. Тоді припустимо, що золото в першій скриньці. А інша скринька ховає в собі сердиту жабу, тоді як скринька з істинним написом — золото, що виливається в істинність і другого твердження. Тепер припустімо, що перший напис хибний і золото в першій скриньці. Тоді другий напис був би...»

Король наказав кинути блазня у підземелля.

Наступного дня блазня привели до короля у кайданках і показали йому дві скриньки.

«В одній скриньці ключ, — почав король, — що звільняє тебе від кайданів; знайдеш ключ — свобода твоя. Та якщо програєш, в іншій скриньці лежить кинджал для твого серця».

На першій скриньці написано було:

«Або обидва написи істинні, або обидва — хибні».

На другій ішлося:

«У цій скриньці лежить ключ».

Блазень міркував так: «Нехай перший напис істинний. Тоді й другий теж мусить бути істинним. Тепер припустімо, що й перший напис хибний. Тоді, знову ж, другий напис мусить бути істинним. Якщо перший напис істинний і також якщо перший напис хибний, у другій скриньці мусить бути ключ. Отже, логічно, що ключ у другій скриньці».

Блазень відкрив другу скриньку і знайшов кинджал.

«Як?! — прокричав блазень у страхові, поки його відтягали. — Логічно неможливо!»

«Цілком можливо, — відказав король. — Я просто наніс ті написи на двох скриньках і поклав кинджал у другу».

(За мотивами твору Ре́ймонда Сму́лліана.)

Притча про болúголов

> «Усі люди смертні. Сократ — людина. Отже, Сократ смертний». — Арістотель(?)

Сократ підніс до губ чашу із болиголовом... — Думаєте, що навіть настоянка болиголова не зможе вбити такого мудрого і доброчесного чоловіка? — запитав один із глядачів. — Ні, — відповів інший глядач, студент філософського факультету, — усі люди смертні, а Сократ — людина; а якщо смертний вип'є болиголову, він має померти. — А що як Сократ *не* смертний? — Дурниці, — відрізав студент, — усі люди смертні *за визначенням*; це ж частина того, що ми закладаємо у слово «людина». Усі люди смертні; Сократ — людина; а отже, Сократ смертний. Це не якась там здогадка, а *логічна істина*. — Може й так... — кинув глядач. — О, дивись: поки ми говорили, Сократ уже випив настоянку. — Так, він може звалитися з ніг будь-якої хвилини, — сказав студент. *І вони чекали-чекали, чекали-чекали...* — Не видно, щоб Сократ був смертним, — констатував глядач. — Тоді Сократ, має бути, не людина, — відповів студент. — Усі люди смертні; Сократ не смертний; тож Сократ не людина. І це не якась здогадка, а *логічна істина*.

Основоположна проблема ствердження, що щось є істинним «за визначенням», полягає у тому, що, вибравши інше визначення, ви не зможете змінити реальність[1].

Можливо, ви роздумуєте так: «За моїми спостереженнями, усі ті істоти, які носять одяг, спілкуються мовою і використовують інструменти, мають і інші спільні риси, як-от дихання і циркуляція крові. Останні тридцять «людей» із цієї сукупності, за якими я спостерігав, коли вони пили настоянку болиголова, невдовзі впали і перестали рухатися. Сократ носить тогу, вільно говорить давньогрецькою, а ще випив настоянку з чаші. Тому я передбачаю, що Сократ звалиться з ніг уже за п'ять хвилин».

1. https://machall.com/view.php?date=2003-04-21

Але це був би лише *здогад*, а не, знаєте, щось абсолютно й навічно певне. Давньогрецькі філософи — як і більшість донаукових філософів — дуже полюбляли певність.

На щастя, у них є і нищівна відповідь на ваші сумніви. Кажуть, що ви неправильно зрозуміли значення твердження «Усі люди смертні». То не якесь там *спостереження*, а частина *визначення* слова «людина». Смертність — одна з кількох властивостей, кожна з яких окремо необхідна, а разом з іншими достатня для визнання належності до класу «людина». Твердження «Усі люди смертні» — логічно чинна і повністю беззаперечна істина. А якщо Сократ є людиною, він *мусить* бути й смертним — ця дедукція настільки беззастережна, наскільки це можливо.

Але тоді ми ніколи не можемо знати напевне, що Сократ — «людина», поки не побачимо його смерті. Нам нічого не дає те, що Сократ вільно говорить давньогрецькою чи що у його жилах тече кров, чи навіть те, що у нього людська ДНК. Жодна з цих характеристик не *тотожна* смертності в *логічному* сенсі. Щоб виснувати його людськість, ви маєте *побачити, як він помре*.

(Та й навіть тоді ви не зможете бути нескінченно певними. А що як Сократ встане з могили після того, як ви стали свідком його смерті? Чи що як він згодився на кріоніку, що більш реалістично? Якщо смертність визначається як скінченна тривалість життя, то вам насправді ніколи не *знати*, чи хтось був людиною, поки ви не станете свідком кінця вічності — просто щоб впевнитися, що вони не вернуться з мертвих. Або ви могли б *подумати*, що ваше спостереження, як Сократ звалюється з ніг, було ілюзією, накладеною на ваші очі за допомогою сканера сітківки. Або ж вам усе це просто примарилося...)

Силогізми проблематичні тим, що вони *завжди* чинні. Твердження «Усі люди смертні; Сократ — людина; отже, Сократ смертний», якщо ви розумієте його як логічний силогізм, чинний у межах нашого всесвіту. Воно логічно чинне і всередині сусідніх гілок Еверетта, в яких через невеликі зміни в еволюційній біохімії болиголов — смачний трунок, а не отрута. Чинне воно і

в усесвітах, в яких Сократа ніколи не існувало чи, як на те пішло, не існувало людей.

За мірками баєсіанства свідчення, яке підкріплює гіпотезу, є таким свідченням, яке ймовірніше спостерігати за умови, що гіпотеза істинна, аніж хибна. Спостереження за чинністю силогізму не може бути свідченням, що підкріплює бодай якесь емпіричне припущення, оскільки силогізм завжди буде логічно чинним, незалежно від істинності чи хибності того припущення.

Силогізми чинні в усіх можливих світах, тому, спостерігаючи за їхньою чинністю, ми ніколи не можемо сказати, у *котрому* з тих можливих світів ми живемо.

Не означає, що від логіки немає користі, — просто вона може сказати нам лише те, що ми *у певному сенсі* й так знаємо. Але ми не завжди віримо тому, що знаємо. Число 29 384 209 просте? На підставі того, як я визначаю свою десяткову систему та аксіоми арифметики, я вже визначив відповідь на це питання, хоч я й досі не знаю відповіді на нього. Щоб знайти відповідь, мені потрібна логіка.

Схожим чином, якщо я формулюю непевне емпіричне узагальнення «Люди вразливі до болиголова» і непевний емпіричний здогад «Сократ — людина», логіка підказує мені, що з попередніх припущень я передбачаю, що і Сократ буде вразливим до болиголова.

Висловлювалася й думка, що логічне міркування можна розглядати як спосіб усунення нашої невизначеності щодо неможливих можливих світів — вилучення ймовірнісної маси з логічно неможливих світів, про логічну неможливість яких ми раніше не знали. У такому сенсі ми можемо розглядати логічний аргумент як спостереження.

Та коли ви говорите про емпіричне передбачення типу «Сократ от-от звалиться з ніг і перестане дихати» чи «Сократ збирається робити п'ятдесят стрибків з розведенням рук, а наступного року — взяти участь в Олімпійських іграх», ви говорите про можливі світи, а не неможливі можливі світи.

Завдяки логіці ми можемо сказати, які гіпотези відповідають яким спостереженням і які з цих гіпотез можуть передбачити майбутнє — вона дозволяє застосовувати старі спостереження і попередні здогадки для розв'язання нової проблеми. Та логіка ніколи не заявляє прямолінійно: «Сократ зараз *перестане* дихати». Логіка ніколи не визначає емпіричні питання, не дає відповіді на жодне питання дійсного світу, яка навіть за найсміливіших припущень може бути будь-якою.

Не забувайте літанії проти логіки:

> Куди б ви не пішли, логіка лишається істинною, Тож вона так і не скаже, де ви стоїте.

Слова — приховані умовиводи

Скажімо, я знайшов бочку, закриту зверху кришкою з отвором, розмір якого дає пролізти руці. Опускаю руку і намацую маленький заокруглений об'єкт. Дістаю і бачу, що він синій — як синє яйце. Знову опускаю руку і намацую щось тверде рівне і гранчасте — дістаю і бачу червоний куб. Так я витягую одинадцять яєць і вісім кубів, і всі яйця сині, а куби червоні.

Опускаю руку і намацую ще один об'єкт форми яйця. Перш ніж дістати й подивитися, я маю зробити припущення. Який він матиме вигляд?

Отримані свідчення не доводять, що кожне яйце у бочці синє, а кожен куб червоний. Свідчення і не те щоб сильні: зрештою, дев'ятнадцять об'єктів — не така вже й велика вибірка. А втім, я припущу, що цей яйцеподібний об'єкт синій (друге найбільш імовірне припущення — червоний). Якщо я поставлю на щось інше, можливостей стільки, скільки розрізнюваних кольорів. Та й хто сказав, що яйце має бути одного відтінку? Може, на ньому взагалі буде зображений кінь.

Тож я кажу: «Синій», але все ще досить скромно ставлюся до своєї відповіді. Я ж людина з отим мудрованим раціоналістським складом розуму — тобто я відстежую свої припущення та уявлення про залежності — принаймні так мені здається, але я усвідомлюю, що це лише припущення... так же?

Але коли з тіні на мене вистрибує здоровезний жовто-смугастий котячий об'єкт, я думаю: «Аааа, тигр!». Я не думаю: «Хм... об'єкти з властивостями величини, жовтизни, смугастості часто вже володіли властивостями голоду і небезпеки, а тому, хоча не обов'язково логічно, це може бути емпірично хороша здогадка, що *ааааааааааа* ХРУСТЬ ХРУСТЬ КОВТЬ».

З якоїсь дивної причини людський мозок, схоже, пристосувався до того, щоб приходити до цього висновку швидко, автоматично і без свідомого відстеження припущень.

І якщо яйцеподібні об'єкти я називатиму «синяйцями», а червоні куби — «червокубами», тоді, просунувши руку у бочку і намацавши ще один яйцеподібний предмет, я скоріше подумаю: «О, синяйце», аніж подумаю про якусь там проблему індукції.

Якщо ви вважаєте, що можете визначити слово, як вам заманеться, то це поширена помилка.

Це було б так, *якби* мозок сприймав слова як чисто логічні утворення, арістотелівські класи, з яких ви берете інформації не більше, ніж закладаєте.

А все ж мозок продовжує свою роботу з категоризації, незалежно від того, схвалюємо ми її свідомо чи ні. «Усі люди смертні. Сократ — людина. Отже, Сократ смертний» — так говорили давньогрецькі філософи. Що ж, якщо смертність — частина вашого логічного визначення «людини», ви не можете віднести Сократа до логічної категорії «людина», доки не побачите його мертвим. Але Арістотель чудово знав, що Сократ — людина, в цьому, власне, й проблема. Мозок Арістотеля відніс Сократа до категорії «людина» так само швидко, як ваш мозок класифікує тигрів, яблука і всю решту середовища — миттєво, тихо і без свідомого затвердження.

Арістотель заклав правила, за якими ніхто не міг виснувати приналежність Сократа до категорії «людина», доки він не помре. А втім, Арістотель зі своїми учнями не переставали доходити висновку, що живі люди — це люди, а тому смертні; вони бачили розрізнювальні властивості, як-от людські обличчя і тіла, а їхні мізки робили стрибок до умовивідної властивості — смертності.

Від хибного розуміння роботи вашого розуму він, на щастя, *не* перестає належно функціонувати. Інакше послідовники Арістотеля померли б з голоду, не змігши виснувати, що об'єкт їстівний, лише тому, що він виглядав і на дотик був як банан.

Тож арістотеліанці і далі класифікували об'єкти навколишнього середовища на основі неповної інформації, як це і робили люди до них. Учні, навчені на арістотелевій логіці, способу

мислення не змінили — зате склали помилкове уявлення про те, *чим* займаються.

Якби ви запитали якогось філософа-арістотеліанця, чи бакалійниця Ке́рол смертна, він відповів би: «Так». Запитаєте, як він дізнався, — почуєте: «Усі люди смертні. Керол — людина. Отже, Керол смертна». Запитаєте, чи це його здогадка чи беззаперечна істина — він обере друге (принаймні, якби питали до XVI століття). Запитаєте, як дізнався, що людина смертна, — відповість, що встановив це за допомогою визначення.

Арістотеліанці все ще були тими самими людьми — вони не втратили свою первісну природу, але сформували хибні переконання про власне функціонування. Вони подивилися у дзеркало самоусвідомлення і побачили щось несхоже на їхнє справжнє «я» — неправильно себе відобразили.

Ваш мозок не бере слова за логічні визначення без емпіричних наслідків, та й ви не повинні. Сам акт словотвору може спонукати ваш розум виокремити категорію і тим самим запустити несвідомий процес висновування на основі схожості. Або заблокувати цей процес; якщо я створюю два ярлики, я можу змусити ваш розум виокремити дві категорії. Помітили, як я сказав «ви» і «ваш мозок», ніби це окремі сутності?

Якщо ви помиляєтеся у тому, що всередині вашої голови, це не змінює її вмісту; інакше Арістотель помер би, якби виснував, що мозок призначений для охолодження крові. Зазвичай філософські помилки не заважають швидким чуттєвим умовиводам.

Але ті філософські помилки можуть серйозно порушити процеси ретельного міркування, за допомогою яких ми коригуємо свої перші враження. Якщо ви вірите, що можете «визначити слово так, як вам заманеться», не усвідомлюючи, що ваш мозок продовжить класифікувати без вашого свідомого контролю, тоді ви не докладатимете зусиль для мудрого підбору визначень.

Обсяг та зміст поняття

> — Що таке червоний? — Червоний — це колір. — Що таке колір? — Колір — це властивість об'єктів.

А що таке об'єкт? А властивість? Невдовзі ці двоє загубляться у лабіринті слів, визначених іншими словами. Цю проблему свого часу описав[1] Стівен Ха́рнед як вивчення китайської за китайсько-китайським словником.

Або якби ви запитали мене: «Що таке червоний?», я б вказав на знак «Проїзд без зупинки заборонено», потім на когось у червоній футболці, на червоний сигнал світлофора. Показав би на кров, що виступила з пальця через мою необережність із ножем, на червону візитівку, а потім відкрив би на комп'ютері колірне коло й навів курсор на червону ділянку. Можливо, цього б і вистачило, хоча справді суворі вчителі — якщо ви розумієте значення слова «ні» — наполягали б, щоб я показав на небо і сказав «ні».

Гадаю, я вкрав цей приклад у Се́мюеля Хаяка́ви (хоч і не певен: чув його колись давно, у туманному дитинстві). (Коли мені було дванадцять, батько випадково видалив усі мої файли на комп'ютері, тож я нічого не пам'ятаю до цього віку.)

Так, пам'ятаю, вперше засвоїв різницю між інтенсіональним та екстенсіональним визначенням[2]. Дати «змістове визначення» означає витлумачити слово чи словосполучення за допомогою інших слів — як у словниках. Дати ж «обсягове визначення» означає вказати на зразки, приклади — як роблять дорослі, коли навчають дітей. Попереднє речення дає змістове

1. https://www.southampton.ac.uk/~harnad/Papers/Harnad/harnad90.sgproblem.html

2. Надалі використовуватимуться поняття «зміст» (інтенсія) та «обсяг» (екстенсія). У сполученні з іменниками — відповідні прикметники «змісто́вий» та «обсяго́вий». — Прим. пер.

визначення «обсягового визначення», що робить його обсяговим прикладом «змістового визначення».

У голлівудських стрічках та інших творах масової культури «раціоналістів» зображують як людей, які схибилися на словах і ширяють у нескінченному вербальному просторі, відірваному від реальності.

Але справжні традиційні раціоналісти давно вже наполягають на тісному зв'язку з досвідом:

> «Якщо ви зазирнете у підручник з хімії за визначенням літію, ви, певно, прочитаєте, що це елемент з атомною масою близько 7. Але якщо в автора логічний склад розуму, він розкаже інакше: серед мінералів, схожих на скло — напівпрозорих, сірих або білих, дуже твердих, крихких і нерозчинних — є такий, що забарвлює безколірне полум'я в червонуватий відтінок. Якщо подрібнити цей мінерал разом із вапном або вітеритом (щурячою отрутою) і перетопити, він частково розчиниться у соляній кислоті. Якщо цей розчин випарувати, залишок обробити сірчаною кислотою і належно очистити, звичайними методами його можна перетворити на хлорид. А якщо цей хлорид у твердому стані перетопити й піддати електролізу за допомогою пів дюжини потужних елементів, вийде кулька рожево-сріблястого металу, що плаватиме на бензині. Отриманий матеріал є зразком літію».[3] — Чарльз Се́ндерс Пірс

Перед вами приклад «логічного розуму», як його описує справжній традиційний раціоналіст, а не голлівудський сценарист.

Але зауважте: Пірс не показує вам *фізичний* шматок літію. Він не прикріпляв його до сторінки своєї книги. Він радше дає мапу скарбів — змістово визначену процедуру, після виконання

3. Peirce, C. S. (1932). Collected papers of Charles Sanders Peirce: Vol. 2. Elements of logic (C. Hartshorne & P. Weiss, Eds.). Harvard University Press. § 4, para. 330 (https://isidore.co/misc/Res%20pro%20Deo/Peirce/Done/2.html#:~:text=the%20Individual%20named.-,330,-.%20If%20you%20look)

якої ви отримаєте об'єкт, котрий він зараховує до літію. Вам все ще не дали шматок літію, але й не сказали: «атомна маса 7». (Хоча якщо у вас *досить гострий* зір, фраза «три протони», може, одразу допоможе впізнати літій...)

Тож коли ви хочете пояснити комусь, що вкладаєте у якесь поняття, змістове та обсягове *визначення* у вашому розпорядженні. Те, що я виклав вище про «визначення», стосувалося способу *передати* поняття — *пояснити*, що ви маєте на увазі під «червоним», «тигром», «людиною» чи «літієм». Тепер поговорімо про поняття як такі.

Справжній зміст мого поняття «тигр» — нейронна схема (у моїй скроневій частці), що аналізує вхідний сигнал із зорової кори й визначає, бачу я тигра чи щось інше.

Справжній обсяг мого поняття «тигр» — усе те, що я називаю тигром.

Змістові визначення не охоплюють усієї сукупності критеріїв розпізнавання, як і обсягові не охоплюють усієї сукупності об'єктів, які я зараховую до визначуваного класу. Якщо я вказую на одного тигра і кажу «тигр», комунікація може провалитися, якщо слухач вирішить, що я маю на увазі «небезпечного звіра», «тигра-самця» чи «жовтий об'єкт». Схожим чином, якщо я скажу «небезпечний жовто-чорний смугастий звір», не вказуючи на щось, слухач може уявити велетенських шершнів.

Словами не охопити всіх тонкощів когнітивного поняття — такого, яким воно існує у вашому розумі, — що дозволяє вам розпізнавати одні об'єкти як тигрів, а інші як нетигрів. У ньому забагато властивостей. Вказати на всіх тигрів, яких коли-небудь бачили, ви не можете, а про все, що *назвали б* тигром, і мови нема.

Найсильніші визначення поєднують змістову й обсягову комунікацію, щоб якомога чіткіше окреслити поняття. І навіть так ви передаєте не самі категорії, а лише *описи* понять або інструкції до їх формування. *Самих* категорій — такими, що вони існують у вашому розумі чи в реальності, — ви не передасте.

(Так, при досить творчому підході можна витворити винятки до цього правила: «Речення, опубліковані Еліезером Юдковські станом на 4 лютого 2008 року, що містять термін "хураґалоні"». Ось я й показав усе, що процедура розпізнавання визнає належним до цього класу. Та за винятком математики, визначення зазвичай являють собою мапи скарбів, а не сам скарб.)

Тож ось вам ще одна причина, чому не можна «визначити слово так, як вам заманеться»: ви не можете напряму завантажувати поняття у мозок іншої людини.

Навіть у рамках арістотелівської парадигми, де ми вдаємо, що визначення тотожні самим поняттям, неможливо *одночасно* довільно задавати і зміст, і обсяг поняття. Припустімо, я тлумачу Марс як «величезну червону скелясту сферу, на 50% віддаленішу від Сонця, маса якої становить приблизно десяту частину маси Землі». А ось показати, що це змістове визначення накладається на певний об'єкт у моєму досвіді чи хоча б який-небудь дійсний об'єкт, — це вже інша справа. А якщо натомість скажу «Це — Марс» і покажу на червону цятку у нічному небі, то довести, що це зовнішнє світло відповідає будь-якому конкретному змістовому визначенню, яке я пропоную, — або моїм змістовим переконанням, як-от «Марс — це бог війни», — вже зовсім інша справа.

Однак більша частина роботи мозку, пов'язаної із застосуванням змісту поняття, відбувається на рівні напівусвідомленого. Ми не усвідомлюємо, що ототожнення червоної цятки з «Марсом» — це окреме питання, відмінне від нашого словесного визначення «Марс — це бог війни». Неважливо, якого штибу змістове визначення я витворю для опису Марса, адже мій розум вірить, що «Марс» позначає цей об'єкт[4]— четверту планету від Сонця.

Коли ви зважаєте на те, як людський розум функціонує на ділі, уявлення «Я можу визначити будь-яке слово так, як мені заманеться» невдовзі переростає в «Я можу вірити про певну

4. https://en.wikipedia.org/wiki/File:Mars_Hubble.jpg

множину об'єктів що завгодно» або «Я можу довільно змінювати критерій належності так, щоб будь-який об'єкт проходив або не проходив цю перевірку». Так само, як словами не передати весь зміст поняття, бо за ним стоїть складний нейронний тест на належність, ви не можете довільно *контролювати* цей зміст, бо він застосовується напівусвідомлено. Ось тому так люблять говорити, що XYZ істинне «за визначенням». Якби визначення змінювалися подібно до емпірично порожніх операцій, як вони й мали б, ніхто б не завдавав собі клопоту сперечатися про них. Та щойно ви хоч трохи зловживаєте визначеннями, вони перетворюються на чарівну паличку (правда, у суперечках — не в реальності).

Скупчення подібностей

Одного разу філософи з Академії Платона заявили, що найкраще визначення людини — «двонога істота без пір'я». Кажуть, Діоген Сінóпський (відомий ще як Діоген Кінік) тут же приніс обскубану курку і заявив: «Ось вона — платонівська людина». Платоністи ж відразу змінили своє визначення на «двонога істота без пір'я з широкими нігтями».

Жоден словник чи енциклопедія на світі ще не перелічував усіх спільних для людей ознак. У нас червона кров, по п'ять пальців на кожній руці, череп з кісток, двадцять три пари хромосом — але все це наявне і в інших видах царства тварин. Ми створюємо складні інструменти для виготовлення складних інструментів, використовуємо комбінаторну синтаксичну мову, беремо енергію з критичних реакцій ядерного поділу — ці ознаки можуть виокремити людей з-поміж тварин, але не всіх, адже багато хто з нас ніколи не будував ядерний реактор. За допомогою правильного набору необхідних і достатніх послідовностей генів ви можете виокремити усіх людей — і тільки їх (принаймні наразі), — але ті гени все ще не будуть *усім*, що є спільного для людей.

Але поки вам не трапиться обскубана курка, вказівка «Шукай двоногу істоту без пір'я» може допомогти вичленити декілька конкретних об'єктів — людей, — а не будинки, вази, сендвічі, котів, кольори чи математичні теореми.

Щойно визначення «двоноге без пір'я» прив'язане до *конкретних* двоногих без пір'я, ви можете проглянути цю групу і почати збирати *інші* ознаки — крім безпір'яності й двоногості, — які «двоногі без пір'я», вочевидь, поділяють також. Ті конкретні двоногі без пір'я, яких ви бачите щодня, ще послуговуються мовою, створюють складні інструменти, говорять комбінаторною мовою із синтаксисом, у них тече червона кров (якщо їх штрикнути), а ще вони вмирають, коли випивають болúголову.

Ось так категорія «людина» збагачується новими ознаками. І коли Діоген зрештою приносить обскубану курку, ми не ведемося на його витівку: вона, очевидно, не схожа на інших «двоногих без пір'я».

(Якби арістотелівська логіка добре моделювала людську психологію, платоністи подивилися б на обскубану курку і сказали: «Так, це людина. До чого ти ведеш?».)

Якщо *перше* двоноге без пір'я, яке вам випало побачити, — обскубана курка, тоді ви можете прийти до висновку, що вербальний ярлик «людина» позначає обскубану курку; тож я можу внести такі зміни до своєї мапи скарбів, щоб вона вказувала на «двоногих без пір'я і з широкими нігтями», а потім, якщо я мудрий, то додам: «Бачиш он там Діогена? От він людина, я людина, ти людина; шимпанзе не людина, хоч і досить близький до неї».

Початкова підказка має лише привести користувача до скупчення подібностей — групи об'єктів, об'єднаних багатьма ознаками. Опісля підказка виконала своє завдання, і я можу далі передавати нову інформацію «наразі люди смертні» чи щось іще, що я хочу сказати про нас — двоногих без пір'я.

Тому словник краще уявляти собі не як книгу визначень арістотелівських класів, а як книгу підказок для зіставлення словесних позначень зі скупченнями подібностей або з властивостями, корисними для розрізнення скупчень подібностей.

Типовість та асиметрична подібність

Птахи вміють літати. Але не страус. Котрий із птахів більш типовий — вільша́нка чи страус?

Котре з крісел більш типове — робоче, крісло-гойдалка чи безкаркасне?

Більшість скаже, що вільшанка — більш типовий птах, а робоче крісло — більш типове крісло. Когнітивні психологи, котрі досліджують такого роду явища експериментально, називають його «ефектом типовості» або «ефектом прототипу» (Rosch and Lloyd 1978)[1]. Наприклад, якщо ви попросите піддослідних позначити твердження типу «Вільшанка — це птах» чи «Пінгвін — це птах» як «істинні» або «хибні», час реакції буде коротшим для більш центральних прикладів. (Я все ще розпаковую свої книги, але досить певен, що вичитав про це у книзі Ле́йкоффа 1986 року[2].) Показники типовості добре корелюють при використанні різних дослідницьких методів: час реакції — один із прикладів; можна навіть просити людей на шкалі від одного до десяти оцінити, наскільки приклад (конкретна вільшанка) накладається на категорію («птах»).

У нас є ментальний показник типовості (можливо, він виконує роль евристики). Але чи існує відповідна похибка, за допомогою якої ми можемо встановити його?

Котре із цих тверджень здається вам більш природним: «дев'яносто вісім приблизно дорівнює ста» чи «сто приблизно дорівнює дев'яносто восьми»? Для більшості у першому

1. Rosch, Eleanor and B. B. Lloyd, eds. (1978). Cognition and Categorization. Hillsdale, N.J.: Lawrence Erlbaum Associates.

2. Lakoff, George. (1986). *Women, Fire and Dangerous Things: What Categories Tell Us About the Nature of Thought*. University of Chicago Press, Chicago.

твердженні було більше сенсу (Sadock 1977)[3]. З тих же причин люди, яких попросили оцінити подібність Мексики до Сполучених Штатів, систематично ставили вищі оцінки за тих, хто оцінював, наскільки Сполучені Штати подібні до Мексики (Tversky and Gati 1978)[4].

І якщо вам здається, що в цьому немає нічого такого, то дослідження Ріпса 1975 року показало, що люди частіше очікували, що хвороба на острові поширюється від вільшанок до качок, аніж від качок до вільшанок. Звичайно, це не *логічно* неможливо, але практично — що б там не відрізняло качку від вільшанки та зменшувало ймовірність передачі хвороби від качки до вільшанки, мусить бути різницею між вільшанкою і качкою і зменшувати ймовірність передачі хвороби від вільшанки до качки.

Звісно, можна придумати собі й такі раціоналізації: «Можливо, поблизу водиться більше споріднених із вільшанками видів, через що хвороба на початку поширювалася б легше, абощо», але дивіться: не перестарайтеся із раціоналізуванням оцінок ймовірностей від піддослідних, які навіть не усвідомлювали, що відбувається порівняння. І не забудьте, що Мексика більше схожа на Штати, ніж Штати на Мексику, а дев'яносто вісім ближче до ста, ніж сто до дев'яносто восьми. Це явище просто пояснити тим, що люди використовують (продемонстровану) евристику схожості як приблизний показник імовірності поширення хвороби, і ця евристика є (вочевидь) асиметричною.

Канзас напрочуд близький до центру Сполучених Штатів, тоді як Аляска напрочуд віддалена від нього; тож Канзас, імовірно, ближчий до більшості міст у Штатах, тоді як Аляска віддаленіша. Однак з цього не слідує, що Канзас ближчий до Аляски, ніж Аляска до Канзасу. Та виглядає так, що люди мислять

3. Sadock, Jerrold. (1977). "Truth and Approximations." In Papers from the *Third Annual Meeting of the Berkeley Linguistics Society*, pp. 430-39. Berkeley: Berkeley Linguistics Society.

4. Tversky, Amos and Itamar Gati. (1978). "Studies of Similarity". In Rosch and Lloyd (1978).

(метафорично) близькість як внутрішню властивість Канзасу, а дальність як внутрішню властивість Аляски; тому й Канзас усе ще близький (навіть відносно Аляски), тоді як Аляска все ще віддалена (навіть відносно Канзасу).

Отже, знову ми бачимо, що арістотелівське уявлення про категорії — логічні класи, належність до яких визначається набором властивостей, кожна з яких строго необхідна, а всі разом строго достатні — погано моделює людську когнітивну психологію. (Хто б міг подумати, що погляд науки дещо змінився за останні 2350 років?) Ми навіть не мислимо так, ніби належність до множини є властивістю зі значенням «істина» або «хиба»: твердження про належність до множини можуть бути більш чи менш істинними. (Зауважте: *не* те саме, що більш чи менш імовірними.)

Ось вам ще одна причина не вдавати, що ви (чи хтось інший) *дійсно* будете розглядати слова як арістотелівські логічні класи.

Структура скупчень у просторі можливих об'єктів

Поняття конфігураційного простору позначає спосіб, яким *описи* об'єктів перетворюють у його *координати*. Може здаватися, що синій «ближчий» до синьо-зеленого, ніж до червоного. Але наскільки? На це питання складно відповісти, якщо просто розглядати кольори. Але корисно знати, що в моделі RGB (пропорційні) координати кольорів становлять 0:0:5, 0:3:2 та 5:0:0. А якщо нанести їх на 3D-графік, стане ще ясніше.

Так само ви бачите вільшанку такою, якою бачите: коричневий хвіст, червона груднина, стандартна форма вільшанки, максимальна швидкість польоту (без вантажу), типова для її виду ДНК та індивідуальні алелі. Або ви можете побачити вільшанку як точку у конфігураційному просторі, виміри якого описують усе, що ми знали чи могли б знати про неї.

Вільшанка більша за вірус і менша за авіаносець — так її можна описати за виміром «об'єм». Схожим чином можна описати її і за виміром «маса»: важить більше за атом Гідрогену і менше за галактику. У різних вільшанок значення вимірів «об'єм» і «маса» сильно корелюватимуть — їхні точки даних вишикуються у майже рівну пряму. Та кореляція не буде досконалою, тому нам все ще потрібні два окремі виміри.

В цьому і перевага погляду на вільшанок як на точки у просторі. Так просто ви б не побачили цей лінійний ряд, якби уявляли вільшанок як милих крихіток, що тріпотять крильцями.

Їхня ДНК — надзвичайно багатовимірна змінна, але її все одно можна вважати частиною положення вільшанки у просторі об'єктів — мільйони координат із чотирма можливими значеннями, по одній на кожну нуклеотидну основу ДНК — або, можливо, у більш витонченому представленні... Форму вільшанки і її колір (відбивна здатність поверхні) ви так само

можете розмістити у просторі об'єктів як частину її положення, навіть якщо їх не передати *одним* виміром.

Так само, як точка координат 0:0:5 містить ту саму інформацію, що й код HTML для блакитного кольору, ми не втрачаємо інформацію, коли розглядаємо вільшанок як точки у просторі. Незалежно від того, чи ми уявляємо фізичну сімдесятиграмову вільшанку, яку зважують на вагах, чи точку даних із масовою координатою +70, ми віримо у те саме твердження про масу вільшанки.

Можна навіть уявити собі конфігураційний простір з одним або кількома вимірами для кожної окремої ознаки об'єкта, так що *положення* точки, що представляє об'єкт, у цьому просторі відповідатиме *всій* інформації, що міститься в самому неабстрактному об'єкті. Щоправда, таке подання було б дещо надлишковим: серед координат одночасно фігурували б маса, об'єм і щільність.

Думаєте, завеликий простір? Квантові фізики застосовують конфігураційний простір із *нескінченною кількістю вимірів*, а одна точка у тому просторі описує положення кожної частинки у всесвіті. Тож ми ще відносно консервативні у нашому представленні простору можливих об'єктів: точка у цьому *просторі* описує лише один об'єкт, а не цілий всесвіт.

Якщо ми не певні точної маси й об'єму вільшанки, то можемо уявити в просторі об'єктів невелику хмарку — *об'єм невизначеності*, у межах якого може лежати її справжнє положення. Щільність цієї хмарки відображає нашу суб'єктивну ймовірність того, що вільшанка має відповідну масу та об'єм. Якщо ви більш певні щільності вільшанки, ніж її маси та об'єму, ваша хмаринка ймовірності буде більш густою довкола виміру щільності та зосереджена навколо похилої лінії у підпросторі маси/об'єму. (Насправді тут хмарка є поверхнею через відношення VD = M.)

Неарістотелівські межі слів когнітивні психологи описують поняттям «радіальна (променева) категорія». Початкова «мати» зачинає дитя, дає йому життя і доглядає. А донорка яйцеклітини, яка ніколи не бачила свого дитяти, може вважатися

матір'ю? Вона «генетична мати». Як щодо жінки, якій пересадили чужорідний ембріон, який вона виношує до пологів? Вона «сурогатна мати». А якщо жінка виховує дитину, яка не її генетично? От вона «прийомна мати». Арістотелівський силогізм спрацював би так: «Люди мають десять пальців. У Фреда дев'ять пальців. Отже, Фред не людина», але насправді ми думаємо так: «Люди мають десять пальців. Фред — людина. Отже, Фред — людина із дев'ятьма пальцями».

Променевість (радіальність) категорій можна осмислити у термінах змістовості. З прикладу вище видно, що деякі властивості зазвичай наявні, а деякі можуть бути відсутні. Якщо розглядати внутрішні критерії слова «мати», його можна уявити як розсіяне сяйво, інтенсивність якого відповідає ступеню, в якому цей об'єм простору об'єктів відповідає категорії «мати». Світіння зосереджене в центрі (генетика, народження і виховання); світіння у донорок яйцеклітин також помітне, хоч і не таке яскраве.

Або ж можна розглядати променевість категорій у розрізі множини об'єктів. Скажімо, за допомогою міри відстані, що максимально точно відповідає тому, як люди сприймають подібність, ми нанесли усіх птахів на простір об'єктів. Вільшанка більше схожа на іншу вільшанку, ніж на голуба, проте вільшанки і голуби все ще більше схожі одне на одного, ніж на пінгвіна.

Тоді центр птаховості щільно населяють численні сусідні види: вільшанки, папуги, канарки, голуби та багато інших. Орли, соколи та інші хижі види займають скупчення, наближчені до них. Скупчення пінгвінів, як і курок та страусів, розміститься ще далі.

Усе це виглядатиме як скупчення астрономічних тіл: багато галактик, що обертаються довкола центру, та декілька відхилень.

Або можна розглянути ще й зміст когнітивної категорії «птах», і її обсяг у реальних птахів: центральні скупчення вільшанок і папуг яскраво сяють як надзвичайно типові птахи; супутникові скупчення страусів і пінгвінів світяться дещо тьмяніше

як нетипові птахи; Абрахам Лінкольн за кілька мегапарсеків від них не світиться взагалі.

Мені більше подобається остання ілюстрація — блискучі точки, — оскільки, як на мене, структура когнітивного змісту випливає зі структури обсягового скупчення. Спочатку існувала «структура у світі» — емпіричний розподіл птахів у просторі об'єктів; потім, побачивши її, ми сформували категорію, змістове сяйво якої приблизно накладається на цю структуру.

Ці ілюстрації дають нам ще один погляд на те, чому слова не є арістотелівськими класами: емпірично скупчені структури дійсного всесвіту не настільки кристалізовані. Природне скупчення — група дуже подібних між собою об'єктів — може взагалі *не* мати набору необхідних і достатніх властивостей, тобто такої множини ознак, які мали б усі члени групи і не мав би жоден об'єкт поза нею.

Та навіть якщо категорія безнадійно розмита і хитка, не поспішайте панікувати. Не маю нічого проти того, що хтось каже: птахи — це «пернаті істоти, що літають». *Але ж пінгвіни не вміють літати!* Нічого страшного. Не треба лізти у вмиральну яму через те, що правила зазвичай мають винятки. Від визначень не можна очікувати, що вони бездоганно лягатимуть на емпіричну структуру простору можливих об'єктів кожного разу, тому що мапа менша і простіша за територію. Завдання визначення «пернаті істоти, що літають» — привести слухача до скупчення птахів, а не дати вичерпний опис кожного існуючого птаха аж до молекулярного рівня.

Коли ви проводите межу довкола групи обсягових точок, *емпірично* скупчених у просторі об'єктів, ви можете знайти принаймні один виняток від кожного простого змістового правила, створеного вами.

Проте якщо визначення достатньо добре працює на практиці, коли треба вказати на задумане емпіричне скупчення, заперечення проти нього можна справедливо назвати прискіпуванням до дрібниць.

Замасковані запити

Уявіть себе за незвичною роботою на незвичній фабриці: ви берете об'єкти з таємничої конвеєрної стрічки і розкладаєте їх по двох контейнерах. На вашій першій зміні старша сортувальниця Сюзен пояснює, що сині яйцеподібні об'єкти називають «синяйцями» і йдуть у «контейнер для синяєць», а червоні куби називають «червокубами», які кладуть у «контейнер для червокубів».

Узявшись до роботи, ви помічаєте, що синяйця і червокуби різняться між собою ще деякими ознаками, крім кольору і форми. На поверхні синяєць росте хутро, тоді як червокуби гладенькі. Синяйця злегка пружні на дотик, тоді як червокуби тверді. Синяйця матові, а поверхня червокуба злегка просвічує.

Невдовзі ви помічаєте синяйце незвично темнішого відтінку, а якщо придивитися, то й колір уже більше нагадує пурпуровий — щось між червоним і синім.

Стривайте! Чому ж тоді ви називаєте цей об'єкт «синяйцем»? Від початку «синяйце» визначили як синій яйцеподібний об'єкт — якість си́ні фігурує у самому слові. А цей об'єкт не синій. Бракує однієї з необхідних якостей — отже, слід називати його «пурпуровим яйцеподібним об'єктом», а не «синяйцем».

Але виявляється: крім пурпуровості та схожості на яйце, воно ще й вкрите хутром, еластичне і матове. Тож, побачивши його, ви подумали: «О, синяйце дивного кольору». Це ж точно не червокуб... так же?

І все ж ви не певні, що з ним робити. Гукаєте старшу сортувальницю Сюзен.

— О, так, це синяйце, — каже вона, — можеш покласти у відповідний контейнер. Рука вже тягнеться до того контейнера, але кидати пурпурове синяйце не поспішаєте: — Сюзен, а як ти *знаєш*, що це синяйце? Старша сортувальниця кидає на вас здивований погляд: — А хіба це не

очевидно? Цей об'єкт може бути пурпуровим, та він усе ще форми яйця, пухнастий, еластичний і матовий, як і решта синяєць. Треба бути готовим до того, що можуть з'явитися декілька порушень кольору. Чи ти говориш про одну з тих філософських головоломок «Як ви знаєте, що світ не було створено п'ять хвилин тому, а вам не завантажили повністю хибні спогади?». У філософському сенсі я не *на всі сто певна*, що це синяйце, але виглядає як хороша здогадка. — Ні, я про те... — ви підшукуєте вдалі слова. — *Чому* взагалі існують окремі контейнери для синяєць і червокубів? Яка між ними *різниця*? — Синяйця синього кольору і форми яйця; червокуби червоного кольору і форми куба, — терпляче відповідає Сюзен. — Тобі ж проводили стандартну ознайомчу лекцію? — Чому синяйця і червокуби *потрібно* сортувати? — Емммм... бо інакше вони всі перемішаються? — каже Сюзен. — Робота сама себе не зробить. Якби ми не сиділи тут цілий день і *не* відділяли синяйця від червокубів, ніхто б нам не платив. — Хто від початку визначив, що перший синій яйцеподібний об'єкт буде «синяйцем»? І як він це визначив? Сюзен стенула плечима: — Думаю, ти міг би так само називати червоні куби «синяйцями», а сині яйця «червокубами». Але так, здається, легше запам'ятати. Ви задумуєтеся: — Припустімо, з конвеєра впав надзвичайно химерний об'єкт. Наприклад, помаранчевий сфероподібний пухнастий напівпрозорий об'єкт зі звивистими зеленими мацаками. Як тоді сказати: червокуб це чи синяйце? — Ого, *настільки* химерної штуковини не знаходив ще ніхто, — відповідає Сюзен, — але в такому випадку ми б поклали його до сортувального сканера. — А як працює цей ваш сортувальний сканер? — питаєте ви. — Рентгенівське випромінювання? Магнітно-резонансна томографія? Спектроскопія пропускання швидких нейтронів? — Мені казали, що воно працює на теоремі Баєса, але я не до кінця розумію як, —

відповідає Сюзен. — Та мені подобається, як воно звучить. Баєс Баєс Баєс Баєс Баєс. — І *яку інформацію дає* тобі сортувальний сканер? — Інформацію про те, в який контейнер класти об'єкт — для червокубів чи синяєць. Ось тому його й називають сортувальним сканером. На цьому у вас закінчилися питання. — Раптом тобі цікаво, — починає Сюзен, — синяйця містять у собі дрібні самородки ванадієвої руди, а червокуби — шматочки паладію. Що те, що те має промислову цінність. — Знущаєшся? — Завжди будь ласка.

Нарешті ми збагнули суть синяйцевості: синяйце — це ємність для самородків ванадієвої руди. Зовнішні ознаки — синій колір і хутро — не *визначають* синяйцевість об'єкта; зовнішні ознаки мають значення лише тому, що вони допомагають *виснувати*, чи є у цьому об'єкті ванадій.

Вміст ванадію — необхідне і достатнє визначення: усі синяйця зберігають ванадій, а все, що зберігає ванадій, є синяйцем: «синяйце» — таке умовне позначення «об'єкта із ванадієм всередині». Так же?

«Не так швидко, — каже Сюзен, — ванадій містять близько 98% синяєць, натомість паладій — 2%. Точніше, — продовжує вона, — близько 98% синіх яйцеподібних пухнастих еластичних матових об'єктів ховають у собі ванадій. Для незвичних синяєць цей відсоток може змінюватися: ванадій містять 95% пурпурових синяєць і 92% твердих синяєць тощо».

Тепер припустімо, що вам трапляється синій яйцеподібний пухнастий еластичний матовий об'єкт — звичайнісіньке собі синяйце — і заради цікавості ви відносите його до сортувального сканера, котрий видає: «паладій» — ті рідкісні 2%. То воно синяйце?

Перше бажання — назвати його «червокубом», позаяк ви збираєтеся кинути його до контейнера з червокубами. Однак виявляється, що при вимкненому світлі майже всі синяйця ледь світяться у темряві, тоді як майже всі червокуби не видають

ніякого світіння у тих же умовах. А відсоток синяєць, що світяться у темряві, не сильно різниться від синіх яйцеподібних пухнастих еластичних матових об'єктів із паладієм всередині замість ванадію. Отже, якщо ви захочете вгадати, чи об'єкт сяє як синяйце чи зовсім не світиться як червокуб, то маєте ставити на перше.

То воно *насправді* червокуб чи синяйце?

З однієї сторони, ви кинете об'єкт у контейнер для червокубів, хай що там іще дізнаєтеся. З іншої сторони, якщо існують ще якісь невідомі ознаки, які вам потрібно виснувати, ви прийдете до них, виходячи з того, що перед вами не червокуб, а синяйце — себто віднесете його у скупчення синіх яйцеподібних пухнастих матових штуковин, а не червоних кубоподібних гладеньких твердих напівпрозорих об'єктів.

Питання «Чи цей об'єкт синяйце?» може означати різні запити в різних ситуаціях.

Якби воно взагалі не відповідало на хоча б *якийсь* запит, вам не було б до того діла.

Атеїзм — це «релігія»? Трансгуманізм — це «секта»? Ті, хто стверджують, що атеїзм є релігією, «бо він заявляє про переконання щодо Бога», насправді намагаються говорити (як на мене), що методи міркування в ужитку атеїзму прирівнюються до релігійних методів міркування чи що атеїзм небезпечніший за релігію у плані ймовірності спричинення насилля тощо... Насправді на кону стоїть претензія атеїстів на суттєву відмінність і зверхність щодо релігії, яку вірянин намагається відкинути, заперечуючи не вищість, а відмінність(!).

Та найбільш абсурдна частина у тому, що у ході суперечки хтось дістає словник і шукає визначення «атеїзму» та «релігії». (Так, це однаково абсурдно — робить це атеїст чи вірянин.) Як словник *взагалі може* визначити, чи емпіричне скупчення атеїстів дійсно суттєво відрізняється від емпіричного скупчення теологів? Як реальність може залежати від значення слова? По тому, як ви перемалюєте межі, точки в емпіричному розподілі даних не посунуться.

Та люди часто не *усвідомлюють*, що їхня суперечка про проведення межі визначення насправді є суперечкою про висновування ознáки, поширеної серед більшості об'єктів всередині емпіричного скупчення...

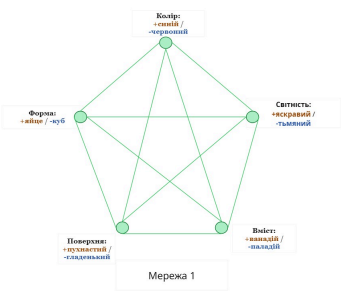
Звідси й словосполучення «замаскований запит».

Нейронні категорії

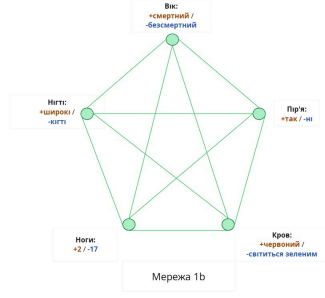
В есеї «Приховані запити» я писав про задачу класифікації «синяєць» і «червокубів». Типове синяйце синє, форми яйця, пухнасте, еластичне, матове, світне у темряві та містить ванадій. Типовий червокуб червоний, форми куба, гладенький, твердий, просвітний, але не світиться в темряві, містить паладій. З міркувань простоти пропоную забути про такі якості, як еластичність/твердість та матовість/просвітність. Так у нас лишається п'ять вимірів у просторі можливих об'єктів: колір, форма, поверхня, світність та вміст.

І от я хочу створити штучну нейронну мережу (ШНМ), що зі спостережуваних ознак синяйця передбачуватиме неспостережувані. Виходимо з того, що моє уявлення про ШНМ вельми наївне: в одній науково-популярній книзі я вичитав, наскільки нейронні мережі розподілені, емерджентні та паралельні — *просто як мозок людини (!!)*. Але в мене ніяк не виходить вивести диференціальні рівняння для градієнтного спуску у нециклічній багатошаровій мережі із сигмоїдними елементами (що набагато легше, ніж здається).

Потім я накреслю нейронну мережу десь такого вигляду:



Мережа 1



Мережа 1b

Мережа 1 призначена для класифікування синяєць та червокубів. Але оскільки «синяйце» — незнайоме і штучне поняття, для порівняння я додав схожу Мережу 1b, призначену для розрізнення людей від космічних чудовиськ: у неї на вході визначення Арістотеля («Усі люди смертні») та Академії Платона («Двонога істота без пір'я із широкими нігтями»).

Нейронна мережа потребує правила навчання. Очевидно, що якщо два вузли часто активні одночасно, зв'язок між ними слід посилювати — це одне з перших запропонованих правил тренування нейромережі, відоме як правило Хе́бба.

Так, якби ви бачили об'єкти і синього кольору, і пухнасті — одночасно активували б вузол «колір» у стані + і вузол «поверхня» у стані +, — зв'язок між ними посилився б: колір+ активує поверхню+ і навпаки. Спостережувані об'єкти комбінації синього кольору, форми яйця і ванадію посилювали б позитивні взаємні зв'язки між кольором, формою та вмістом.

Скажімо, за час роботи на конвеєрі ви вже побачили чимало синяєць і червокубів. Але тепер вам трапилося щось пухнасте, яйцеподібне та — ойой! — червонувато-пурпурове (змоделюємо у «колір» з рівнем активації -2/3). А ви ж іще навіть не тестували світність чи вміст. Що передбачати? Що передбачати?

Внаслідок цього рівні активності в Мережі 1 дещо коливаються. Позитивна активація перетікає з форми у світність, негативна — з кольору у вміст, негативна — зі світності у вміст... Певна річ, усі ці команди передаються *паралельно (!!)* та *асинхронно (!!)* — зовсім як у людському мозку...

Зрештою, Мережа 1 стабілізується і проявляє високу позитивну активацію для вимірів «світності» та «вмісту». Можна сказати, що мережа «очікує» (хоча ще не побачила), що об'єкт буде світитися у темряві і містить ванадій.

Та ось Мережа 1 проявляє таку поведінку, навіть якщо не існує явного вузла, що свідчить: синяйце перед вами чи ні. Це судження *неявно присутнє по всій мережі (!!)*. Синяйцевість — це *атрактор (!!)*, котрий постає внаслідок *емерджентної поведінки (!!)* з *розподіленого (!!)* правила навчання.

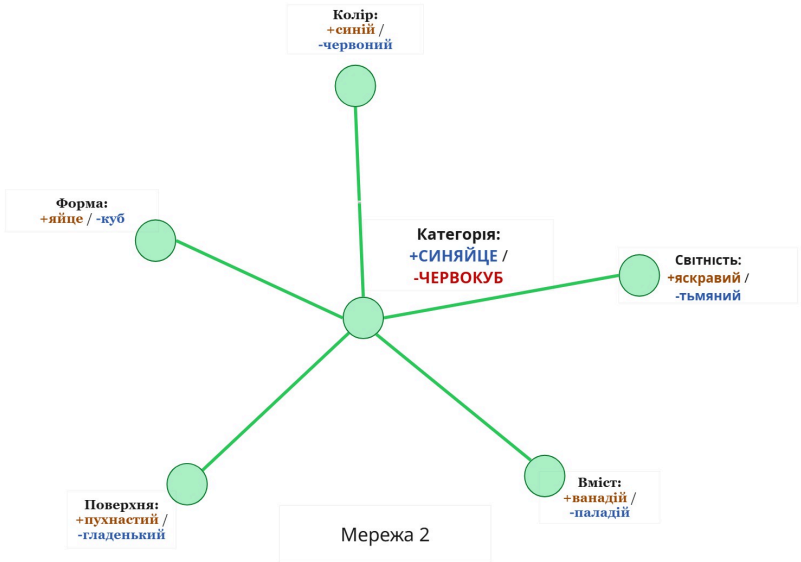
А в реальному житті такого роду мережева архітектура — як банально[1] це не звучало б — стикається з *цілою низкою* проблем. Циклічні мережі не завжди стабілізуються відразу: вони можуть коливатися, демонструвати хаотичну поведінку чи просто довго стабілізуватися. Ситуація, коли ви бачите щось велике, жовте і посмуговане і маєте ще почекати п'ять хвилин, щоб ваша розподілена нейронна мережа стабілізувалася до атрактору «тигр», вийде вам боком. Нехай це й асинхронно та паралельно, але не в режимі реального часу.

Крім того, вистачає й інших проблем, серед яких подвійне врахування свідчень, коли інформація передається туди й назад: підозрюєте, що об'єкт яскраво світиться в темряві, — активується переконання про вміст ванадію, а воно, своєю чергою, знову активує переконання, що об'єкт світиться в темряві.

А якщо спробуєте масштабувати архітектуру Мережі 1, для цього знадобиться $O(N^2)$ зв'язків, де N — загальне число спостережуваних величин.

Тож якою могла б виглядати більш реалістична архітектура нейромережі?

1. https://thedailywtf.com/Articles/No,_We_Need_a_Neural_Network.aspx



У цій мережі хвиля активації прямує від усіх зафіксованих (спостережуваних) вузлів до центрального вузла, а потім знову поширюється від нього до всіх незафіксованих (неспостережуваних) вузлів. Отже, ми можемо обчислити відповідь за один крок, а не чекати, поки мережа стабілізується — важлива вимога в біології, де нейрони активуються лише на частоті 20 Гц. А архітектура мережі масштабується як O(N), а не $O(N^2)$.

Варто визнати, що деякі нюанси легше помітити в архітектурі першої мережі, ніж другої. Мережа 1 має прямий зв'язок між кожними двома вузлами. Тож якщо червоні об'єкти *ніколи* не світяться у темряві, але червоні пухнасті об'єкти мають інші ознаки синяєць — форма яйця і ванадій, — Мережа 1 може легко відобразити такий стан речей: потрібен дуже сильний прямий негативний зв'язок з кольору до світності, але сильніші позитивні зв'язки з поверхні до решти вузлів, за винятком світності.

І те, що червоні об'єкти не світяться, не якийсь «особливий виняток» із загального правила про світність синяєць. У Мережі 1 немає ж частин, які представляють синяйцевість — вона виникає як атрактор у розподіленій мережі.

Тому так — ті зв'язки N2 давали нам якусь користь. Та й не дуже велику. У більшості реальних задач Мережа 1 не надто *корисніша* від другої, адже у природі вкрай рідко можна знайти звіра, застряглого десь посередині між котом і собакою.

(Крім того, є ще й дані, які не так просто відобразити у Мережі 1 *чи* Мережі 2. Скажімо, сполучення морського відтінку синього і форми сфери завжди позначає наявність паладію. Але коли вони ці ознаки знаходять окремо (як у виключній диз'юнкції), кожна з них слугує дуже сильним свідченням наявності ванадію. Без додаткових вузлів це складно відобразити у будь-якій архітектурі. Що Мережа 1, що Мережа 2 втілюють невисловлені припущення про те, якого роду структура середовища більш імовірна. У машинному навчанні саме здатність розпізнавати такі припущення відрізняє фахівців від новачків.)

Не переплутайте: ані Мережа 1, ані Мережа 2 не є біологічно реалістичними. *Але* все ж виправданим здається припустити, що робота мозку, якою вона б не була, ближча до Мережі 2 ніж до першої. Принцип «швидко, дешево, масштабовано» добре працює для розрізнення собак і котів: природний добір діє в таких випадках, як вода, що стрімко тече по ландшафту пристосованості.

Здавалося б, що такого надзвичайного у завданні розсортувати об'єкти на синяйця чи червокуби по відповідних контейнерах. А чи помітили б ви, якби об'єкти морського відтінку не світилися в темряві?

Можливо, якби хтось показав вам двадцять об'єктів, схожих між собою лише цим кольором, а потім вимкнув світло, жоден з них не засвітився б. Іншими словами, якби вас ткнули носом. Певно, якщо показати вам усі ті об'єкти кольору моря, ваш мозок сформує нову підкатегорію і зможе виявити ознаку «не світиться» у цій же підкатегорії. Але ви, певно, так і не помітите їх,

якщо ті об'єкти будуть розкидані серед сотень синяєць і червокубів. Уже не так *легко* чи *інтуїтивно*, як із розрізненням котів і собак.

А от із «Сократ — людина. Усі люди смертні. Отже, Сократ смертний». Звідки Арістотель знав, що Сократ — людина? Так, у Сократа не росло пір'я, були широкі нігті, він ходив на двох, говорив давньогрецькою, мав людську тілобудову і поводився як людина. Тож мозок, зрештою, доходить висновку, що Сократ — людина, і на цій підставі припускає, що він смертний, як і всі інші спостережувані люди. Запитати, наскільки носіння одягу, на відміну від користування мовою, асоціюється зі смертністю, не виглядає простим чи інтуїтивним завданням. Просто «істоти, які носять одяг і користуються мовою, — це люди», а «люди смертні».

Чи виникають біаси, пов'язані зі спробою віднести об'єкт до категорії раз і назавжди? Певна річ. Про це можете прочитати в есеї «Сектантське сектоборство».

Далі буде...

Як алгоритм відчувається зсередини

«Чи видає звук дерево, яке падає у безлюдному лісі?» — пам'ятаю, як із цього приводу почалася цілком наївна суперечка, далека навіть від берклiанського суб'єктивного ідеалізму. От як тут:

> — Воно видає звук, як і інші дерева при падінні! — Але який може бути звук, якщо його ніхто не чує?

За типовим раціоналістським поглядом, перший говорить за «звук» у значенні акустичних коливань у повітрі, а другий — у значенні слухового досвіду. Відповіді на питання «Чи існують акустичні коливання?» та «Чи існують слухові досвіди?» очевидні без роздумів. Тож насправді суперечка йде за визначення слова «звук».

Як на мене, стандартний аналіз у цілому коректний. Візьмімо його за основу і запитаймо: чому люди беруть участь у такого роду суперечках? Яка психологія стоїть за ними?

Основна ідея програми евристик і біасів полягає в тому, що помилки часто дають більше уявлення про процес мислення ніж правильні відповіді. Вступати у палку суперечку про те, чи при падінні дерево видає звук, традиційно вважається помилкою.

Тож якого роду будова розуму відповідає цій помилці?

В есеї «Приховані запити» я впровадив класифікаційне завдання на розрізнення синяєць та червокубів. Старша сортувальниця Сюзен вводила вас у суть роботи: сортувати об'єкти з конвеєрної стрічки треба по контейнерах — сині яйця мають лежати у контейнері для синяєць, а червоні куби — у контейнері для червокубів. Виявляється, такий розподіл зумовлюється наявністю дрібних самородків ванадієвої руди у синяйцях, червокуби ж містять паладій. Обидві копалини важливі для промисловості.

За винятком близько 2% синіх яйцеподібних об'єктів, всередині яких паладій. То якщо ви знаходите синій яйцеподібний об'єкт із паладієм, його треба називати «червокубом»? Ви ж все одно покладете його у контейнер з червокубами. Чому б не назвати його «червокубом»?

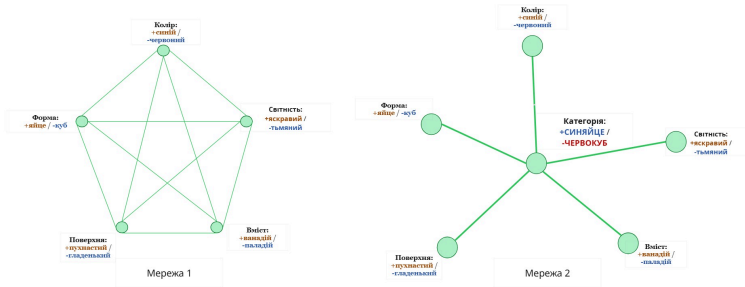
А коли вимикаєте світло, майже усі синяйця ледь світяться в темряві. І сині яйцеподібні об'єкти з паладієм світитимуться в темряві з такою ж імовірністю, як і будь-який інший синій яйцеподібний об'єкт.

Якщо ви знаходите синій яйцеподібний об'єкт із паладієм усередині й питаєте: «Це синяйце?», відповідь залежить від того, що ви робитимете з нею. Якщо питаєте «У котрий контейнер іде об'єкт?», тоді обираєте так, ніби у вас червокуб. Але якщо ви питаєте: «Чи сятиме він у вимкненому світлі?», ви діємо так, ніби у вас синяйце. У першому випадку питання «Це синяйце?» позначає прихований запит «У котрий контейнер його треба кидати?», у другому — запит «Чи сятиме він у вимкненому світлі?».

А тепер уявіть синій яйцеподібний об'єкт із паладієм усередині: ви вже пересвідчилися, що на ньому хутро; він гнучкий, матовий і сяє у темряві.

У вас уже є всі *відповіді* на кожен запит, є всі введені спостереження. Замаскованих запитів уже *не лишилося*.

Чому тоді хтось усе ще поривається розвинути суперечку про те, чи об'єкт *дійсно* є синяйцем?



Ця діаграма з есею «Нейронні категорії» показує дві різні нейромережі, за допомогою яких можна відповідати на питання про синяйця та червокуби. У Мережі 1 низка недоліків: потенційно нестабільна/хаотична поведінка чи потреба в $O(N^2)$ зв'язках. Проте структура Мережі 1 має значну перевагу над Мережею 2: кожен мережевий елемент відповідає тестовному запиту. Якщо спостерігати всі спостережувані величини й зафіксувати значення кожної з них, у мережі не залишиться жодного обчислювального елемента.

Однак Мережа 2 куди краще підходить до чогось настільки ж невизначеного, як робота людського мозку: вона швидка, дешева, масштабована і має додатковий підвісний елемент у центрі, який може або активуватися, або ні, навіть після того, як ми спостерегли кожен з навколишніх вузлів.

Тобто навіть по тому, як ви дізналися, якого кольору об'єкт (синього чи червоного), якої форми (яйце чи куб), поверхні (пухнастий чи гладенький), світності (яскравий чи безсвітний) та якого вмісту (ванадій чи паладій), не покидає *відчуття*, що лишилося питання без відповіді — «*Це справді синяйце?*».

У нашому повсякденні акустичні коливання і слуховий досвід ідуть одним комплектом. Проте дерево, що падає у безлюдному лісі, ламає це звичне уявлення. І навіть після того, як при падінні дерево створює акустичні коливання (замість слухового

досвіду), залишається *відчуття*, що питання «*Чи видало воно звук?*» зосталося без відповіді.

Ми знаємо розташування Плутона, як він рухається, знаємо його форму і масу. Але чи є він планетою?

Не забувайте: коли ви дивитеся на Мережу 2, як я її виклав тут, ви бачите алгоритм ззовні. Люди не замислюються штибу «Чи має загорятися центральний елемент?», як і не думають: «Чи має у моїй зоровій корі загорятися нейрон №12 234 320 242?».

Уявити власний мозок ззовні — значне зусилля, і навіть тоді ви не побачите свій справжній мозок. Уявите те, що, *на вашу думку* (сподіваюся, вона ґрунтуватиметься на науці), там є, а втім, зсередини у вас немає прямого доступу до структур вашої нейромережі. Ось тому елліни і не винайшли обчислювальної нейронауки.

При погляді на Мережу 2 ви дивитеся *ззовні*; те, як структура нейромережі відчувається *зсередини* (якщо ви і *є* тим мозком, що запускає алгоритм), провокує питання «То це синяйце чи як?» навіть після того, як ви знаєте кожну ознаку об'єкта.

Така прірва — серйозна перешкода, і я бачив, як вона спиняла людей на пів шляху. Бо інстинктивно свої інтуїтивні відчуття ми розпізнаємо не як «інтуїції», а як безпосередній дійсний світ. Дивлячись на зелену чашку, ви мислите її не як картинку, відтворену у вашій зоровій корі — хоч саме *це* ви і бачите, — ви думаєте, що просто бачите зелену чашку. Ви думаєте: «О, зелена чашка», а не «Картинка у моїй зоровій корі цієї чашки зелена».

І так само, коли люди сперечаються про те, чи дерево при падінні видає звук, чи Плутон належить до планет, вони не бачать, що сперечаються про те, чи слід активувати певну категоризацію у своїх нейромережах. Здається, що дерево або видає звук, або ні.

Ми знаємо розташування Плутона, як він рухається, знаємо його форму і масу. Але чи є він планетою? Аякже, були й люди, які казали: та це бійка за визначення. Але навіть такий погляд походить з Мережі 2, адже так ви сперечаєтеся про те, як саме

треба під'єднати центральний елемент. Якби ваш мозок сконструювали за архітектурою Мережі 1, ви б сказали не «Залежить від того, що ви маєте під "планетою"», а «З того, що ми знаємо про орбіту, форму і масу Плутона, тут немає чого питати». Чи радше так воно *відчувалося* б — як ситуація, коли усі питання вичерпано (якби ваш мозок працював за принципами Мережі 1).

Перед тим, як ставити під сумнів свої інтуїції, маєте усвідомити, що об'єкт вашого внутрішнього взору *є* інтуїцією — якийсь мисленнєвий алгоритм, на який ви дивитеся зсередини, — а не безпосереднім сприйняттям того, якими предмети є насправді.

Гадаю, люди чіпляються за свої інтуїції не тому, що вірять у безпомильну надійність своїх мисленнєвих алгоритмів, а тому, що не можуть подивитися на свої інтуїції *зсередини своїх мисленнєвих алгоритмів*.

Ось так усе, що ви намагаєтеся сказати про помильність власного мисленнєвого алгоритму, обертається, зрештою, у протиставлення їхньому безпосередньому сприйняттю речей, якими вони є насправді, й відкидається як явно помилкове.

Суперечка про визначення

Я не раз спостерігав, як розмови — навіть нібито про когнітивістику, — перетворювалися на суперечки про визначення. На прикладі класичного спору «Якщо дерево падає у безлюдному лісі, то чи видає воно звук?» спір переростає у щось таке:

> *Якщо дерево падає у безлюдному лісі, то чи видає воно звук?*
>
> Альберт: Певна річ. Що за дурне питання? Щоразу, коли я чув, як падає дерево, воно видавало звук. Відповідно, решта дерев при падінні так само видає звуки. Не думаю, що щось зміниться, якщо мене там не буде.
>
> Беррі: Зачекай. А як воно може видавати звук, якщо ніхто не слухає?

Тут Беррі сперечається з Альбертом через різні інтуїції про те, що таке звук. Але стандартний спір може початися багатьма способами. У Беррі міг бути свій привід відкинути Альбертів висновок. Або Беррі міг виступати скептиком, який, почувши Альбертів аргумент, мимоволі почав шукати в ньому можливі логічні хиби. А по тому, знайшовши контраргумент, машинально прийняв його, не спробувавши знайти контраргумент до контраргументу, а відтак раціоналізував свою протилежну позицію. Не обов'язково, що *апріорна* інтуїція Беррі — та, яку він би висловив, якби ми запитали у нього до того, як заговорив Альберт, — відрізняється від Альбертової.

Навіть якби початкова інтуїція Беррі не відрізнялася від його співрозмовника, то тепер ситуація інакша.

> Альберт: Як ти собі таке думаєш — не видає звуку? Коріння тріскається, стовбур із тріском падає і вдаряється об ґрунт. Усе це породжує коливання, які поширюються по

землі та повітрю. Туди ж спрямовується енергія падіння, перетворюючись на тепло і звук. По-твоєму, якщо у лісі немає людей, то дерево порушує закон збереження енергії?

Беррі: Але ж ніхто не чує ніяких звуків. Якщо у лісі анікогісінько і, в рамках аргументу, жодної іншої істоти зі складною нервовою системою, здатною «почути», не лишається нікого, хто б почув звук.

Що Альберт, що Беррі наводять аргументи, які, як їм *відчувається*, підкріплюють власні позиції, докладніше розкриваючи думки, що активували їхні датчики «звуку» чи проігноровані ними. Але досі розмова велася довкола лісу, а не визначень. І зауважте, що вони не розходяться в думках щодо того, що відбувається в лісі.

Альберт: У дурнішому спорі я ще не брав участі. Яке ж ти пиздобалакало.

Беррі: Ось так значить. І це я чую від їбалометного теревенщика.

Образу кинуто — образу прийнято. Тепер жодна зі сторін не може поступитися, адже поступитися — означає втратити лице. Технічно ми вже вийшли за рамки *суперечки*, як її розуміють раціоналісти, але я все одно включаю цю частину, позаяк у стандартній суперечці вона важлива.

Альберт: Дерево видає акустичні коливання. За визначенням, це і є звук.

Беррі: Ніхто його не чує. За визначенням, це не звук.

Ось тепер суперечка починає крутитися довкола визначень. Щоразу, коли, за винятком суперечок про чисту математику, ви відчуваєте спокусу сказати «за визначенням», все істинне «за визначенням» є істинним в усіх можливих світах, тож

спостереження цієї істини ніяк не звужує множину світів, у *яких* ви можете перебувати.

> Альберт: Мікрофон на моєму комп'ютері може записати звук, навіть якщо поруч немає нікого, хто б його почув, і зберегти його у файл, що буде називатися «звукозаписом». І в цьому файлі зберігається саме структура коливань повітря, а не нейронних імпульсів у чиємусь мозку. «Звук» позначає структуру коливань.

Альберт наводить аргумент, який, за його *відчуттями*, підкріплює те, що слово «звук» *означає щось конкретне*. Це питання вже відрізняється від того, чи у лісі справді поширюються акустичні коливання. Проте ця підміна зазвичай проходить непоміченою.

Беррі: Справді? А зазирнімо у словник.

У ситуації із падінням дерева мене може зацікавити багато чого. Я міг би піти у ліс і подивитися на дерева, навчитися виводити хвильові рівняння для змін атмосферного тиску, ознайомитися з анатомією вуха чи опанувати нейроанатомію слухової кори. І замість усього цього, очевидно, треба звернутися до словника. Чому? Чи серед лексикографів є фахівці у ботаніці, фізиці або нейронауці? Можливо, мало б сенс зазирнути в енциклопедію. Але чому саме в *словник*?

> Альберт: Ага! Визначення 2c у словнику Мерріам-Вебстер: «ЗВУК. Механічна енергія випромінювання, що передається поздовжніми коливаннями в матеріальному середовищі (наприклад, у повітрі)».

> Беррі: Ага! Визначення 2b у словнику Мерріам-Вебстер: «ЗВУК. Слухове відчуття.

> Альберт і Беррі (разом): Бісів словник! Жодної помочі!

Укладачі словників фіксують мововжиток, а не регулюють. Вони відшукують слова у нинішньому мовообігу, а поряд записують

(малу частину) того, що люди, вочевидь, мають на увазі під ними. Якщо воно вживається у різних значеннях, укладачі записують більше одного визначення.

Альберт: Дивись: припустімо, я залишив у лісі мікрофон, і при падінні дерева він записав зразки акустичних коливань. Якби я комусь відтворив їх, він би назвав це «звуком»! Так заведено вживати це слово! Не вигадуй своїх химерних визначень!

Беррі: По-перше, я можу дати слову таке визначення, яке захочу, поки застосовую його послідовно. По-друге, визначення, яке я навів, було в словнику. По-третє, а хто дав тобі право вирішувати, що заведено, а що ні?

У цій стандартній суперечці чимало раціоналістських помилок. Деякі з них я вже розписав, а деякі ще маю (як-от когнітивна протидія таким помилкам).

Поки я хочу зауважити — у невтішному сенсі, — що Альберт і Беррі, здається, сходяться в думках з приводу практично всіх процесів, що відбуваються в лісі, і водночас ця згода, видно, не викликає у них відчуття порозуміння.

Суперечки про визначення ведуть кругами: якби люди з самого початку бачили, куди їх заведе ця стежка, то не ходили б нею. Якби ви запитали Альберта чи Беррі, чому він далі сперечається, імовірно, відповідь звучала б так: «Беррі/Альберт намагається підставити своє визначення «звуку» — а грець би йому! — щоб виправдати свою сміховинну позицію. Я прагну відстояти стандартне визначення».

Припустімо, я повернувся назад у часі, ще до початку суперечки:

(Нізвідки у своєрідному транспорті, схожому на машину часу з оригінальної стрічки «Машина часу», з'являється Еліезер.)

Беррі: Хай йому! Мандрівник у часі!

Еліезер: Я прийшов до вас з майбутнього! Послухайте, що скажу вам! Я повернувся до вас у глибоке минуле —

на п'ятнадцять хвилин назад...

Альберт: П'ятнадцять *хвилин*?

Еліезер: ...щоб передати вам повідомлення!

(Западає суміш спантеличення з очікуванням.)

Еліезер: Як ви думаєте: «звук» слід тлумачити як сукупність акустичних коливань (передача хвиль тиску у повітрі) та слухових відчуттів (наявність того, хто сприймає цей звук) чи, навпаки, лише як акустичні коливання або лише як слухові відчуття?

Беррі: Ти повернувся у часі, щоб запитати нас про *це*?

Еліезер: Мої цілі — моя справа! Відповідай!

Альберт: Я... я не розумію, чому це так важливо. Ви можете взяти будь-яке визначення, поки застосовуєте його послідовно.

Беррі: Підкиньте монетку. Еее, двічі.

Еліезер: На мою думку, якщо виникає така проблема, обом сторонам слід перейти до опису події однозначними складниками нижчого рівня, як-от акустичні коливання чи слуховий досвід. Або кожна сторона може створити нове слово — наприклад, «альберзль» та «берґулум», — щоб відповідно позначати те, що вони називають «звуком»; і тоді вже кожна сторона може послуговуватися новими словами послідовно. Так жодній стороні не доведеться поступатися чи втрачати лице, і вони все ще зможуть спілкуватися. І, звісно, треба постійно відстежувати ту перевірювану тезу, довкола якої крутиться спір. Таке рішення вам здається правильним?

Альберт: Певно що...

Беррі: Чому ми говоримо про це?

Еліезер: Щоб зберегти вашу дружбу на випадок непередбачуваних обставин, яких ви вже ніколи не дізнаєтеся. Бо майбутнє вже змінене!

(Еліезер з машиною зникає у хмарі диму.)

Беррі: Нагадай, на чому ми зупинилися.

Альберт: А, так — якщо дерево падає у безлюдному лісі, чи видає воно звук?

Беррі: Видає альберзль, але не берґулум. Яке наступне питання?

Ця протидія не усуває *усіх* суперечок про категоризацію. Але усуває значну їх частину.

Відчути значення

Коли хтось говорить мені: «Гляди! Метелик!», проговорені фонеми [ме^{и}те́ли^{е}к] потрапляють у моє вухо, створюють вібрації у моїй барабанній перетинці, передаються до за́витки, подразнюють слухові нерви, що передають імпульси активації до слухової кори, де починаються процеси обробки фонеми, розпізнавання слів, відтворення синтаксису (ні разу не послідовний процес) та іншого роду складнощі.

Та зрештою (а точніше — наприкінці цієї секунди) я подивлюся, куди показує друг, і побачу візуальний образ, який розпізнаю як метелика. А якщо натомість раптом побачу вовка, то неабияк здивуюся.

Мій друг дивиться на метелика, його горло продукує звукові коливання, губи ворушаться; хвилі тиску непомітно для людського ока поширюються у повітрі, моє вухо вловлює їх, нерви передають, а мозок розшифровує і так далі і тому подібне — і ось я розумію, на що дивиться мій друг. Хіба не фантастика? Яким неймовірним стало б таке відкриття, якби ми не знали про хвильові коливання! Усі газети в один голос: «Люди — телепати! Люди здатні обмінюватися думками!».

Та ми *і так* телепати. Але магія вже не так зачаровує, коли вона всього лише *частина буденної реальності* і всі ваші друзі теж уміють користуватися нею.

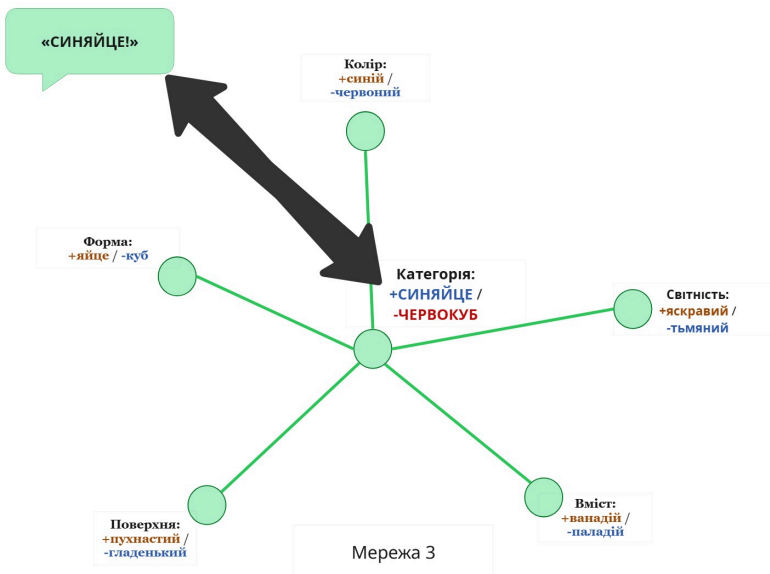
Думаєте, телепатувати просто? А спробуйте скласти комп'ютер, що телепатуватиме з вами. Телепатія, або «мова», або як ви ще назвете здатність частково передавати думки, куди складніша, ніж видається на перший погляд.

Доволі незручно постійно думати: «А тепер я маю частково виразити деякі з моїх думок у лінійну послідовність фонем, які спровокують схожі думки у мого співрозмовника...».

Тому мозок ховає цю складність — а точніше, взагалі ніколи не відображає її, — внаслідок чого люди думають щось дивакувате про слова.

Як я писав раніше, коли смугаста істота стрибає на мене, я думаю: «Аааaa! Тигр!», а не «Хм... істоти з властивостями величини, жовтизни, смугастості часто вже володіли властивостями голоду і небезпеки, а тому, хоча не обов'язково логічно *ааааааааааа* ХРУСТЬ ХРУСТЬ КОВТЬ».

Схожим чином, коли хтось кричить «Аааaa! Тигр!», природний добір не буде прихильний до організму, який думає: «Хм... я почув склад [тигр], який мої одноплемінники асоціюють з внутрішніми аналогіями мого поняття "тигр". Найімовірніше, вони вимовлятимуть його, коли вони категоризують *айайаааааа* ХРУСТЬ ХРУСТЬ *рятуйте — моя рука* ХРУСТЬ КОВТЬ».



Якщо взяти цю особливість за обмеження в архітектурі людського мислення, ви не захочете, щоб між моментом розпізнавання складу [тигр] вашою слуховою корою та моментом активації поняття «тигр» були *якісь* додаткові етапи.

Повертаючись до притчі про синяйця та червокуби та централізованої мережі, яка швидко й недорого класифікує інформацію, ви можете уявити прямий зв'язок від блока, що розпізнає фонеми [сиᶜн′айцé], до центрального мережевого елемента. А центральний блок — поняття «синяйце» — активується майже відразу по тому, як ви чуєте: «Синяйце!» від старшої сортувальниці Сюзен.

До процедури (яка теж не має тривати вічність): щойно ви бачите синій яйцеподібний об'єкт і центральний блок синяйця загоряється, ви кричите Сюзен: «Синяйце!».

А от зсередини алгоритм відчувається так, ніби позначення та поняття майже *ідентичні* — тобто значення *відчувається* як внутрішня невіддільна властивість самого слова.

Знавці розпізнають у цьому ще один випадок «помилки проєкції розуму», описаної Едвіном Джéйнсом. Складається враження, ніби значення слова притаманне самому слову, ніби червоність — властивість самого яблука, а таємничість — властивість таємничого явища.

У більшості випадків мозок взагалі не розрізняє слово від значення, хіба що може завдати собі клопоту розділяти їх, коли вивчає нову мову. І навіть тоді, коли Сюзен покаже на синій яйцеподібний об'єкт і скаже: «Синяйце!», ви подумаєте: «*Цікаво, що означає "синяйце"*», а не «*Цікаво, яку ментальну категорію Сюзен асоціює зі слуховим позначенням "синяйце"*».

А тепер подивіться з цього боку на стандартний спір про визначення, в якому дві сторони сперечаються про те, що *насправді* означає слово «звук», як могли б сперечатися, чи конкретне яблуко *насправді* червоне чи зелене:

> Альберт: Мікрофон на моєму комп'ютері може записати звук, навіть якщо поруч немає нікого, хто б його почув, і зберегти його у файл, що буде називатися «звукозаписом». І в цьому файлі зберігається саме структура коливань повітря, а не нейронних імпульсів у чиємусь мозку. «Звук» позначає структуру коливань.

Беррі: Справді? А зазирнімо у словник.

Альберт інтуїтивно відчуває, що слово «звук» *має конкретне значення* — акустичні коливання. Так само як Альберт відчуває, що при падінні у лісі дерево *видає звук* (а не спричиняє подію, що *належить до категорії звуку*).

Беррі відчуває:

звук.значення == слуховий досвід ліс.звук == хибно

Але не:

мійМозок.ЗнайтиПоняття("звук") == поняття_СлуховеВідчуття поняття_СлуховеВідчуття.віднести(ліс) == хибно

Друге ближче до того, що відбувається *насправді*; але люди не еволюціонували до розуміння цього, як і до інтуїтивного знання, що мозок складається з нейронів.

Конфліктні інтуїції Альберта й Беррі підживлюють подальшу суперечку, коли вона переходить до питання про те, що означає слово «звук». Така суперечка *здається* цілком звичайною емпіричною суперечкою про стан речей — ніби сперечатися про те, чи небо синє чи зелене.

Ви можете навіть не помітити, що збилися з дороги, аж поки не спробуєте виконати раціоналістський ритуал — сформулювати перевірний експеримент, результат якого залежить від тих обставин, про які ви так палко сперечаєтеся...

Аргумент від поширеного вжитку

Під час стандартного спору про визначення ви почуєте щось таке:

> Альберт: Дивись: припустімо, я залишив у лісі мікрофон, і при падінні дерева він записав зразки акустичних коливань. Якби я комусь відтворив їх, він би назвав це «звуком»! Так заведено вживати це слово! Не вигадуй своїх химерних визначень!
>
> Беррі: По-перше, я можу дати слову таке визначення, яке захочу, поки застосовую його послідовно. По-друге, визначення, яке я навів, було в словнику. Третє, а хто дав тобі право вирішувати, що заведено, а що ні?

Не всі суперечки про визначення доходять до етапу, коли згадують поняття «поширений ужиток». Гадаю, найчастіше люди звертаються до словника, бо вірять, що слова містять значення, а словник вірогідно фіксує його. Здається, дехто навіть вірить, що словник *визначає* значення, що лексикографи регулюють мову. Можливо, це переконання пов'язане з тим, що ще за школи їхні авторитарні вчителі казали, що вони мусять керуватися словником, адже це правило не факультативне, а обов'язкове.

Лексикографи ж вивчають тексти письменників і записують те, що, на їхню думку, ці слова означають. Вони історики. Оксфордський словник англійської мови, може, і *вичерпний*, але точно не *авторитетний*.

Але як же без суспільної необхідності застосовувати слова у загальноприйнятому значенні? Чи не ґрунтується робота нашої людської телепатії — цінної мовної сили — на взаємній узгодженості? Може, нам і слід розглядати лексикографів як вищих арбітрів — навіть якщо *вони* вважають себе істориками, — щоб підтримувати ту непомітну співпрацю, від якої залежить мовлення.

Словосполучення «авторитетний словник» майже ніколи не вживають як слід. Авторитетний словник стандартних термінів Інституту інженерів з електротехніки та електроніки (IIEE) — приклад належного вжитку. До цього Інституту входять члени із правом голосу, яким з професійних міркувань необхідна строга узгодженість термінів та визначень, тож такий Словник — чинний нормативний документ, узгоджений у результаті переговорів, який має ту владу, яку, на думку кожного члена, має Інститут.

У повсякденному житті спільна мова зазвичай не постає внаслідок виважених домовленостей, як у випадку з IIEE. Повсякденна ж мова більше подібна до інфекції, позаяк слова винаходять і поширюють через культуру. (Услід за Річардом Докінзом це явище почали називати «мемом». А якщо ви ще не зрозуміли, про що я, то прошу до гугла — ось і ви вже заразилися.)

Проте, як показав приклад зі Словником IIEE, мовна домовленість теж може бути суспільним благом, створеним на засадах співпраці. Якщо ми хочемо обмінюватися думками за допомогою мови — цієї людської телепатії, — нам вигідно позначати схожі поняття *одними й тими самими* словами. Бажано, щоб ці поняття збігалися з тим, як їх розрізняє наш мозок, — навіть якщо ми не маємо спільного інтересу саме до того, *яке* слово позначає яке поняття.

У нас немає жодного очевидного інтересу вживати слово «о́то» на позначення звуку чи слово «звук» — на позначення о́то. Але у нас є взаємний інтерес вживати *одне й те саме* слово, яким би воно не було. (Бажано, щоб найуживаніші слова були короткими, але поки не буду йти в теорію інформації.)

Та попри спільний інтерес, не геть *обов'язково*, щоб ми з вами використовували подібні позначення для *внутрішнього* вжитку, адже це лише для зручності. Якщо я знаю, що для вас «ото» означає звук — тобто пов'язуєте «ото» з поняттям, дуже схожим на те, яке я пов'язую зі словом «звук», — то я можу сказати: «Зіжмакування паперу дає тріскаюче ото». Для цього потрібне додаткове розумове зусилля, але за бажання не бачу проблем.

Так само, якщо ви запитаєте: «Що за тростина у кулі для боулінгу, що падає на підлогу?» — і я знаю, яке поняття *ви* пов'язуєте зі складами [тростúна], тоді я зможу зрозуміти, що ви маєте на увазі. На це може піти трохи мисленнєвих ресурсів і часу, тому що зазвичай я пов'язую «тростину» з іншим поняттям. Але я зрозумію.

Коли люди дійсно *хочуть* порозумітися одне з одним, нас не зупинити! Якщо ми потрапимо на безлюдний острів і не матимемо спільної мови, то візьмемо палиці й почнемо малювати картинки на піску.

Звертаючись до аргументу від поширеного вжитку, Альберт припускає, що мовний договір — суспільне благо, встановлене спільними зусиллями. І все ж Альберт робить це припущення лише для того, щоб риторично звинуватити Беррі у порушенні цього договору і створенні загрози для суспільного блага. Тепер уся ця суперечка про падаюче дерево з ботаніки перейшла у семантику і політику. Тому у відповідь Беррі кидає виклик Альберту за право визначати це слово.

Раціоналіст, достатньо дисциплінований, щоб триматися свого запиту, помітив би, що ця розмова геть вийшла з берегів.

Любий читачу, усе це правда потрібно? Альберт знає, що Беррі має на увазі під «звуком»; і Беррі знає, що під «звуком» має на увазі Альберт. І перший, і другий знають такі словосполучення, як «акустичні коливання» і «слухові відчуття», які вони вже пов'язують із тими самими поняттями і ними можуть недвозначно описувати події в лісі. Якби вони опинилися на безлюдному острові й намагалися порозумітися, вони б *досягли своєї мети*.

Коли обидві сторони *знають*, що *хоче* сказати інша сторона, і обидвоє звинувачують один одного у викривленні «поширеного вжитку», то що б вони не обговорювали, *способу порозумітися так вони не знайдуть*. А в цьому ж і полягає основна перевага поширеного вжитку.

Навіщо сперечатися про значення слова, якщо обидві сторони намагаються відстояти свої погляди? Якщо це лише

конфлікт простору імен, який надмірно роздули, і на кону нічого більше не стоїть, обидвом сторонам треба просто придумати двійко нових слів і вживати їх послідовно.

Утім, класифікування часто функціонує на підґрунті прихованих умовиводів та замаскованих запитів. А атеїзм — «релігія»? Якщо хтось стверджує, що методи мислення, які застосовують в атеїзмі, нарівні з методами мислення в юдаїзмі або що атеїзм так само породжує насильство, як і іслам, тоді він має очевидну аргументативну зацікавленість у тому, щоб звести все це до однієї невиразної сірої маси під назвою «віра».

Або ще сперечаються за об'єднання чорношкірих та білошкірих до класу «люди». У цьому випадку вигадувати два слова буде недоцільно, адже на кону саме та ідея, що не слід проводити моральне розмежування.

Але щойно на кону стає будь-яке емпіричне *чи* моральне твердження, ви більше не можете покликатися на поширений ужиток.

Якщо питання у тому, як об'єднати схожі об'єкти в скупчення для висновування, емпіричні прогнози залежатимуть від відповіді на це питання — тобто визначення можуть бути *хибними*. А конфлікт передбачень уже не можна розв'язати опитуванням громадської думки.

Якщо ви хочете знати, чи атеїзм слід відносити до надприродних релігій для якихось конкретних емпіричних висновків, словник вам не допоможе.

Якщо ви хочете знати, чи чорношкірі відносяться до людей, словник вам не допоможе.

Якщо всі вірять у те, що червона цятка у небі — то бог війни Марс, словник визначатиме «Марса» як бога війни. Якщо всі вірять у те, що вогонь вивільняє флогістон, словник тлумачитиме «вогонь» як процес вивільнення флогістону.

Використання слів — свого роду мистецтво; навіть коли визначення буквально ні істинне, ні хибне, часто воно може бути мудрішим чи дурнішим. Словники фіксують історію мововжитку; якщо розцінювати їх як вищих арбітрів значення, воно

прив'язує вас до мудрості минулого, закриваючи доступ до покращення.

Тому подбайте-таки (якщо вам доведеться відійти від мудрості минулого), щоб люди могли зрозуміти, що саме ви намагаєтеся сказати.

Порожні позначення

Подивімося (вкотре) на арістотелівську ідею категорій. Скажімо, існує якийсь об'єкт з властивостями A, B, C, D та E (або принаймні дещо схожий на E).

> Фред: Ти маєш на увазі, що ота штука отам блакитна, округла, пухнаста і... Я: В арістотелевій логіці немає значення, якими саме є ці властивості або як я їх називаю. Тому я використовую літери.

Потім я створюю арістотелівську категорію «зава», що описує ті і тільки ті об'єкти, котрі володіють властивостями A, C та D.

> Я: Об'єкт 1 — зава, B та E. Фред: І він — тобто A — теж синій. Так? Я: Це мається на увазі, коли я кажу «зава». Фред: Я хочу-таки, щоб ти сказав це прямо. Я: Гаразд. Об'єкт 1 — це A, B, зава та E.

Тоді я додаю ще одне слово — йокі, — котре описує усі об'єкти лише з властивостями B та E; а за ним — «зіппо», котре описує усі об'єкти з властивістю E, але не D.

> Я: Об'єкт 1 — зава і йокі, але не зіппо. Фред: Стривай. То він світиться? Тобто він E? Я: Так. З наявної інформації іншої можливості немає. Фред: Якби ти ще проговорив це. Я: Добре: Об'єкт 1 — це A, зава, B, йокі, C, D, E і незіппо. Фред: Ого! І ти можеш усе це визначити на око?

Вражає. Правда ж? Можемо придумати ще кілька нових слів: «боло» — це A, C і йокі; «ман» — це A, C і зіппо; а «мерлакдоніан» — це боло і ман.

Безглуздо і заплутано? Теж так думаю. Замінімо ці позначення такими визначеннями:

> «Зава, B і E» стає [A, C, D], B, E «Боло та A» стає [A, C, [B, E]], A «Мерлакдоніан» стає [A, C, [B, E]], [A, C, [E, ~D]]

Про арістотелівські категорії слід запам'ятати, що [A, C, D] — вся інформація про «заву». І справа не в тому, що я можу легко поміняти позначення, а в тому, що вони взагалі не потрібні, адже арістотелівські класи керуються виключно структурами типу [A, C, D]. Люди називають «завою» одну з цих структур — або підв'язують будь-яке інше позначення — суто для зручності (чи незручності), що анітрохи не змінює арістотелівських законів.

Скажімо, «людину» слід визначати як смертне двоноге без пір'я. Тоді класичний силогізм набуває такого вигляду:

> Усі [смертний, ~пір'я, двоногий] смертні. Сократ — [смертний, ~пір'я, двоногий]. Отже, Сократ смертний.

Процес міркування уже не виглядає таким разючим. Еге ж?

У цьому випадку *ілюзія логічного висновку* постає з позначень, що приховують засновки й створюють враження новизни у висновку. Якщо замінити позначення на визначення, ілюзія розсіюється, оприявнюючи тавтологію емпіричної порожнечі. Ви не можете сказати, що Сократ — [смертний, ~пір'я, двоногий], поки не пересвідчитеся, що він таки смертний.

Як ви могли помітити, я терпіти не можу думки «ви можете визначити слово як завгодно». Ця ідея бере початок в арістотелівському баченні категорій, які працюють за умови, що ви слідуєте їм *чітко* й *неухильно* (що точно не про людей). Арістотель чудово знав, що Сократ — людина, хоч це і не випливало з його правил, та *якби* якась вигадана нелюдська сутність слідувала його правилам, то ніколи б не прийшла до суперечності. Та й нічого особливого вона не з'ясувала б: просто не змогла стверджувати, що Сократ — [смертний, ~пір'я, двоногий], поки не пересвідчиться у його смертності.

Але справа не стільки в тому, що позначення в арістотелівській системі довільні, позаяк вона чудово працює *взагалі без позначень*, скільки в тому, що вона видає точно такий самий потік тавтологій, які хіба що не настільки захопливі. Позначення лише створюють *ілюзію* логічного висновку.

Тому, якщо хочете арістотелівський афоризм, він має звучати так: «Я можу визначати слово як мені завгодно» — і навіть не: «Визначення слова ніколи не має наслідків» — а краще: «Визначенням не потрібні слова».

Затабуюйте це слово

У грі «Табу» (від компанії «Ха́сбро») гравець бере картку із загаданим словом і має пояснити його своїй команді, не згадуючи цього слова і п'яти додаткових слів-асоціацій, перерахованих на картці. Наприклад, ваша команда має відгадати слово «бейсбол», а ви не можете вимовляти такі слова, як «спорт», «бита», «удар», «подача», «база» і, звісно ж, «бейсбол».

Щойно я починаю дивитися на задачу через призму гри, мені одразу спадає на думку таке: «Штучний груповий конфлікт, де ви орудуєте довгим дерев'яним циліндром, щоб вгатити по підкинутому сфероїду, а потім пробігти між чотирма безпечними позиціями». Можливо, за правилами, не найефективніший спосіб передати слово «бейсбол». Відповідно, кращим міг би бути варіант «Те, у що грають янкі». Але у цій загальній навичці *викреслювання слова з пам'яті* я вправлявся роками, хоч і з іншою метою.

У попередньому есеї ми побачили, як заміна термінів на визначення може викрити емпіричну непродуктивність класичного арістотелівського силогізму. Усі люди смертні (а ще, вочевидь, двоногі і без пір'я); Сократ — людина; тому Сократ смертний. Коли ми замість слова «людина» підставляємо його очевидне визначення, розкривається ось таке засадниче міркування:

> Усі [смертний, ~пір'я, двоногий] смертні. Сократ — [смертний, ~пір'я, двоногий]. Отже, Сократ смертний.

Та принцип заміни слів на визначення має куди ширше застосування:

> Альберт: При падінні у безлюдному лісі дерево видає звук. Беррі: При падінні у безлюдному лісі дерево не видає звуку.

Очевидно ж, що якщо один каже «звук», а інший — «незвук», у нас виникає явна суперечність. Але припустімо, що перед тим, як почати суперечку, обидвоє розіменують свої покажчики у точки даних:

> Альберт: Дерево, що падає у безлюдному лісі, відповідає [критерію приналежності: ця подія призводить до акустичних коливань]. Беррі: Дерево, що падає у безлюдному лісі, не відповідає [критерію приналежності: ця подія призводить до слухових відчуттів].

Тепер очевидного конфлікту вже немає — а варто було заборонити собі вживати слово «*звук*». А якби почалася суперечка про словосполучення «акустичні коливання», ми б знову зіграли в табу і сказали: «ти́скова хвиля у матеріальному середовищі». За потреби ми затабуюємо і слово «хвиля», замінивши його на хвильове рівняння. (Затабуюйте «слухове відчуття» і отримаєте «форму сенсорної обробки всередині мозку людини, що на вході приймає лінійний часовий ряд поєднань частот...».)

Припустімо, що Альберт і Беррі мають вже іншу суперечку:

> Альберт: Сократ відповідає поняттю [критерій приналежності: ця людина помре, випивши настоянку болиголова]. Беррі: Сократ відповідає поняттю [критерій приналежності: ця людина не помре, випивши настоянку болиголова].

Тепер між Альбертом і Беррі виникла суттєва розбіжність в очікуваннях: вони по-різному уявляють собі, що станеться, якщо Сократ вип'є болиголова. Але якщо для різних понять вони вживатимуть те саме слово «людина», то можуть і не помітити цієї розбіжності.

Залежно від того, чи ви дивитеся з позиції позначення (Альберт каже «звук», Беррі каже «незвук», тож вони, скоріше, не сходяться в думках) чи з позиції критерію (критерій Альберта — акустичні коливання, критерій Беррі — слухові відчуття),

складається вже зовсім інше уявлення про те, з чим вони погоджуються, а з чим ні.

Зберіть групу *так званих* футуристів і запитайте їх, чи вірять вони у появу штучного інтелекту через тридцять років. Думаю, принаймні половина з них скаже, що так. І якщо не втручатися, вони поручкаються і привітають одне одного з досягнутим консенсусом. Але я пропоную затабувати термін «штучний інтелект» і попросити їх описати, *що* вони очікують побачити, не вживаючи слів «комп'ютер» чи «думати». Гадаю, під невиразним загальновживаним терміном ви помітите значний конфлікт очікувань. Як і під оцим терміном. І прочитайте обов'язково ще 71 визначення інтелекту[1] від Ше́йна Ле́ґґа[2].

Розвіяти ілюзію єдності між релігіями можна шляхом табування слова «Бог», а потім запитати, у що саме вони вірять. Або затабуюйте слово «віра» і запитайте, чому вони вірять у те, в що вірять. Хоча здебільшого вони не спроможуться дати бодай якусь відповідь, тому що в першу чергу вони заявляють про свою віру. А на аудіозаписі неможливо когнітивно наблизити окремий фрагмент.

Коли ви постаєте перед філософськими труднощами, *на першій лінії оборони слід не визначати проблематичні терміни, а зрозуміти, чи можете ви взагалі обійтися без них* (будь-яких їхніх скорочених синонімів). І дивіться: не придумайте якогось нового слова їм на заміну. Описуйте зовнішні спостережувані змінні і внутрішні механізми. Не використовуйте тільки один позначник, яким би він не був.

Альберт каже: у людей є «свобода волі». Бері каже: у людей немає «свободи волі». З цього вже проглядається очевидний

1. https://arxiv.org/pdf/0706.3639

2. Найкориснішими виглядають визначення з кластера дослідників штучного інтелекту, які явно встановлюють три критерії влучного визначення: 1) працює для людини, машини, тварини; 2) перевіряється емпірично через досягнення цілі; 3) не прив'язується до одного завдання чи середовища. Мета ж словникових визначень — дати перелік консенсусних ознак інтелекту, а психологічних — виміряти та диференціювати людей між собою. — Прим. пер.

конфлікт. Багато які філософи порадять їм обом спробувати дати точне визначення того, що вони мають на увазі під «свободою волі», на що у них, певна річ, піде чимало часу. Я б натомість порадив їм описати, що, на їхню думку, такого мають чи не мають люди, заборонивши собі вживати словосполучення «свобода волі». (Якщо хочете спробувати вдома, уникайте також слів «обирати», «діяти», «вирішувати», «визначений», «відповідальний», а також їхніх синонімів.)

Ось такий нестандартний засіб з мого набору інструментів, котрий, на мою скромну думку, *куди* ефективніший за будь-який стандартний. Він і більше зусиль потребує, але так ви хоч отримаєте те, за що платите.

Замініть символ суттю

Що потрібно зробити, щоб побачити «бейсбол» — приклад з попереднього есею — як «штучний груповий конфлікт, де ви орудуєте довгим дерев'яним циліндром, щоб вгатити підкинутий сфероїд, а потім пробігти між чотирма безпечними позиціями»? Яких зусиль потребує раціоналістська версія гри «Табу», мета якої не підібрати синонім, відсутній на картці, а підібрати спосіб описати поняття без стандартного словесного позначення?

Тут знадобиться ваша уява. Ваш внутрішній взір має оприявнити усі дрібниці так, ніби ви бачите їх уперше. Маєте скористатися технікою «Свіжий погляд».

Ота штука — то «бита»? Ні, то довгий заокруглений дерев'яний стрижень, що звужується з одного боку, щоб людина могла легко взяти його і розмахувати ним.

А ота штуковина — то «м'яч»? Ні, то шкіряний сфероїд, зшитий симетричним швом, твердий, хоч і не як метал; його можна взяти в руку й кинути або вдарити дерев'яною палицею, або зловити.

А оті ділянки — то «бази»? Ні, то визначені на гральному полі позиції, до яких чимдуж біжать гравці, оскільки, за штучними правилами гри, там вони в безпеці.

Найбільша завада у погляді на речі свіжими очима — наявність у розумі зручного підсумку, готової поняттєвої опори: «бейсбол», «бита», «база». Зусилля потрібне, щоб не давати розуму скочуватися по напрацьованому шляху, по легкому шляху, по шляху найменшого спротиву, куди заходить невеличке невиразне слово і затьмарює всі деталі, що ви їх так силкуєтеся розгледіти. Саме слово може набувати руйнівної сили кліше і нести в собі отруту закешованої думки.

Табувати слово — могти описати щось без допомоги стандартного вказівника, позначення чи посилання — одна з *основних* компетенцій раціоналіста. Вона на тому ж ґрунтовному

рівні, що й звичка постійно питати «Чому?» або «Чого я можу очікувати від цього переконання?».

Це мистецтво тісно пов'язане з:

- прагматизмом, оскільки таке бачення дає нам тісніший зв'язок з очікуваним досвідом, на відміну від висловленого переконання;
- редукціонізмом, оскільки таке бачення змушує нас перейти на нижчий рівень побудови і звернути увагу на частини замість того, щоб поглядом пробігати до цілого;
- технікою «Триматися свого питання», оскільки слова часто відривають нас від питання, яке ми насправді хочемо поставити;
- відмовою від закешованих думок, що в'їдаються у загальновживані слова, так що табуюванням цих слів ви можете блокувати кеш;
- письменницьким принципом «Показуй — не розказуй!», котрий має силу серед раціоналістів;
- зосередженням на початковій меті.

Як табуювання слова може допомогти вам триматися своєї цілі?

З есею «Втрачені цілі»:

> «Поки ви читаєте есей, якийсь хлопець чи дівчина сидить за університетською партою і старанно вивчає матеріал, який не збирається застосовувати і до якого не має внутрішнього інтересу. Студенти хочуть високооплачувану роботу, яка, своєю чергою, вимагає шматка паперу, для якого потрібен попередній ступінь магістра, а для цього — ступінь бакалавра, а університет, що дає той ступінь, потребує від вас проходження курсів техніки в'язання XII століття. Тож вони ретельно навчаються, щоб після кінцевого іспиту зрештою забути вивчене, але і на

тому не зупиняються, бо ж вони так *хочуть* той папірець».

Навіщо ви ходите до «школи»? Отримати «освіту», що закінчиться «дипломом». Зітріть заборонені слова та їхні очевидні синоніми, уявіть, чим саме займаються у школі, і, скоріше за все, ви помітите, що зараз «школа» обмежується сидінням поряд зі знудженими підлітками і вислуховуванням того, що ви й так уже знаєте, що «диплом» — це папірець з якимись писульками, а «освіта» прирівнюється до забування матеріалу одразу ж після іспиту.

Нещільні узагальнення часто проявляються у вигляді категоризації: тих, хто навчається у класах, відносять до категорії «здобувачі освіти», тож «здобування освіти» має бути чимось правильним; а коли хтось просто з'являється у виші, також підпадає під поняття «здобувач освіти», незалежно від того, навчається цей хтось чи ні.

Учні, які розуміють математику, отримуватимуть хороші оцінки на контрольних, але якщо ви вимагаєте, щоб школи демонстрували хороші результати на контрольних, там витрачатимуть увесь час на підготовку до контрольних. *Ментальна категорія*, що не повністю відповідає вашій меті, може породити аналогічну *внутрішню* втрату мотивації. Ви хочете навчатися — вам потрібна «освіта»; тож поки ви отримуєте щось відповідне категорії «освіта», можете й не помітити, навчаєтеся ви насправді чи ні. Або помітите, але не зрозумієте, що відхилилися від початкової мети, адже ви «здобуваєте освіту», і саме так для себе описали свою мету.

Категоризувати — означає виключати інформацію. От вам кажуть, що падаюче дерево породжує «звук», а ви не знаєте, що то за звук; самі ви не чули, як впало те дерево. Якщо монетка випадає гербом, ви не знаєте її радіальної орієнтації. Синій яйцеподібний об'єкт може бути «синяйцем». Але що як його форма змінюється від яйця до яйця, а синій має різні відтінки? Ви хочете використовувати категорії, щоб виключати недоречну

інформацію, щоб просіювати золото від піску, та часто стандартна категоризація виливається у відсіювання і доречної інформації. А коли ви вже стикаєтеся з такою проблемою у мисленні, найочевидніше рішення — зіграти в табу.

Наприклад: поняття «зіграти в табу» — саме по собі нещільне узагальнення. Версія гри від «Хасбро» відрізняється від раціоналістської версії: у першій на картці перелічують лише п'ять додаткових заборонених слів, але цієї кількості і близько недостатньо, щоб обмежити мислення знайомими старими словами. Те, що роблять раціоналісти, зараховується як гра в табу — відповідає поняттю «грати в табу», — але не все, що рахується за гру в табу, запускає процес свіжого погляду. Подумаєте: «зіграти в табу, щоб по-новому оцінити ситуацію», — вирішите, що усе, зараховане до гри в табу, відповідатиме свіжому погляду.

Раціоналістська версія не є грою, тому виграти хитрістю й обходом правил вам не вдасться. Граючи в табу, ви будете змушені накласти на себе добровільні обмеження, наприклад: не вживати синонімів, яких немає на картці. А ще — не вигадувати нових слів, котрі виконуватимуть ту саму функцію у розумі, що й старі. Ви намагаєтеся збільшити масштаб своєї мапи, а не перейменовувати міста; розіменувати покажчик, а не виділяти новий; спостерігати за подіями у міру їхнього розгортання, а не перефразовувати кліше.

Уявляючи проблему у більших подробицях, ви можете помітити втрачену ціль: що саме ви робите, коли «граєте в табу»? Якій меті служить кожна з частин?

Якщо ви свіжим поглядом оцінюєте свої дії та ситуацію, то й побачите і свої цілі безпосередньо. Змога дивитися свіжим поглядом, ніби вперше в житті, дає вам можливість робити те, про що ви раніше й не задумувалися, якби не мали звички.

Мета губиться тоді, коли символ (диплом, оцінка з контрольної, медичне обслуговування) витісняє суть (навчання, знання, здоров'я). Щоб віднайти втрачену мету — чи то пак нещільну категоризацію, — потрібно вдатися до протилежного.

Замініть символ на суть. Замініть позначення на позначуване. Замініть властивість на критерій приналежності. Замініть слово на значення. Замініть позначення на поняття. Замініть узагальнення на подробиці. Замініть опосередковане питання на безпосереднє. Розіменуйте покажчик. Спустіться на нижчий рівень побудови. Змоделюйте процес у себе в голові, а не називайте його. Збільште масштаб на мапі.

«Просту істину» я написав як вправу з опису «істини» на нижчому рівні побудови, уникаючи таких слів, як «точний», «правильний», «представляти», «відображати», «семантичний», «вірити», «знання», «мапа» чи «дійсний». (Не забувайте: я *не* ставив за мету зіграти в табу. У тексті слово «істинний» фігурує, але *не* для визначення істини. У грі такий хід викликав би обурення, але ми *не* граємо у неї. Запитайте себе, чи документ виконав своє призначення, а не наскільки відповідав правилам.)

Сама Баєсова теорема описує «свідчення» чистою математикою — там немає слів «натякає», «означає», «підтверджує», «доводить» чи «виправдовує». Спробуйте *дати визначення* цим філософським термінам і приречете себе ходити колами.

А тоді залишиться найважливіше слово, яке слід затабувати. Я часто попереджав[1] не зловживати ним чи навіть уникати цього поняття у певних випадках. Тепер ви розумієте чому. Непогано було б подумати про це. Адже справжнє розуміння вимірюється вашою здатністю описати, що ви робите і чому, *без* цього слова чи будь-якого з його синонімів.

1. В есеї «Два коани про секти». - Прим. пер.

Помилки стиснення

От кажуть: «Мапа не територія». Тому єдина атомарно деталізована стовідсотково коректна мапа Каліфорнії у натуральну величину — це сама Каліфорнія. Але Каліфорнія має ряд важливих повторюваних особливостей: розмір автомагістралей, що можна описати меншою кількістю інформації (а що вже казати за *матеріал*) порівняно з описом кожнісінького атома в межах кордонів штату. А відтак кажуть *і так*: «Мапа не територія, але територію ви не складете і не покладете собі в кишеню».

На паперовій мапі Каліфорнії масштабом десять кілометрів до одного сантиметра (один до мільйона) просто бракує місця, щоб показати вам точне розташування двох листочків, опалих на узбіччі за сантиметр один від одного. Навіть якби картографи постаралися відобразити листочки, ті виглядали б як суцільна точка на мапі (чи то пак розмір цього об'єкта мав би становити десять нанометрів, а це більша роздільна здатність, ніж та, з якою працює більшість друкарень, а для людського ока і поготів).

Дійсність разюче масштабна: навіть та частина, яку ми можемо бачити, — мільярди світлових років завширшки. А ваша мапа цієї дійсності записана на кілограмі з лишком нейронів, що вмістилися всередині вашого черепа. Не беріть близько до серця, та ваш череп порівняно крихітний.

Неминуче, що деякі відмінні на ділі об'єкти стиснуться в одну точку.

Але зсередини не відчувається так, ніби ви кажете: «А зараз я стисну два об'єкти в один». Зсередини *відчувається* так, ніби існує *один* об'єкт, і ви дивитеся на нього.

Достатньо мала дитина чи достатньо давній грек-філософ не знають про такі явища, як «акустичні коливання» чи «слухові відчуття». А тому падіння дерева у них супроводжується лише однією подією — «звуком».

По суті, усвідомити відмінність цих *двох* подій, яку на вашій мапі позначено *однією* точкою, — справді значне і складне *наукове* завдання.

Іноді помилки стиснення постають внаслідок об'єднання двох знайомих об'єктів під одним позначенням: ви знаєте про акустичні коливання та слухову обробку, і обидва процеси називаєте «звуком», чим збиваєте себе з пантелику. Але якщо ви *взагалі не знаєте*, що *існують* дві окремі сутності, виникає ще загрозливіша помилка. У системі файлів є лише одна інтелектуальна тека з назвою «звук», і все, що стосується «звуку», потрапляє туди. І навіть не дві теки з однаковими назвами — лише одна. Мапа за замовчуванням стиснена. Для чого ж мозку створювати два інтелектуальні контейнери, якщо достатньо й одного?

Або уявіть собі детективний роман, в якому слідчий здогадується, що в одного з підозрюваних є родич-близнюк. У процесі розслідування звичайна робота слідчого зводиться до спостереження: Керол носить червоне, має чорне волосся, її сандалі пошиті зі шкіри — але усі ці *дані стосуються* лише Керол. Поставити під сумнів окремий випадок не складно, от як: НоситьЧервоне(Керол) чи ЧорнеВолосся(Керол). Можливо, ЧорнеВолосся(Керол) хибне. Може, вона фарбує волосся. Може, тоді КоричневеВолосся(Керол). Але лише кмітливий детектив задумається, чи не варто розділити файл «Керол» — той самий файл для його спостережень — із НоситьЧервоне(Керол) і ЧорнеВолосся(Керол) на *два* окремі. Може, є дві Керол, тож Керол, яка носила червоне, не та сама Керол, у якої чорне волосся.

У цьому випадку сам акт *створення* двох різних контейнерів — геніальний проблиск інтуїції. Легше поставити під сумнів свої спостереження, ніж свою онтологію.

Мапа дійсності у людському мозку, на відміну від паперової мапи Каліфорнії, може динамічно розширюватися, коли ми записуємо в неї більш деталізовані описи. Але зсередини воно відчувається не як збільшення масштабу, а як розщеплення

неподільного атома, наче ми беремо *щось єдине* (так воно сприймається) і розщеплюємо на дві чи більше частин.

Часто цей процес проявляється у вигляді створення нових слів, як-от «акустичні коливання» та «слухові відчуття» замість одного «звуку». Створювати нову назву — ніби відводити новий контейнер. Щойно слідчий усвідомлює, що перед ним дві Керол, одну з підозрюваних практично відразу він почне називати «Керол-2» або «Інша Керол».

Але розширювати мапу не завжди так легко, як давати містам нові назви. Усвідомити, що акустичні коливання чи слухові відчуття взагалі *існують*, — як зловити проблиск наукової інтуїції.

Першими дотичними прикладами на думку спадають слова «інтелект» і «свідомість». З кожної праски у медіа нам розповідають про те, що якесь дослідження «пояснило свідомість»: група неврологів дослідила електричний ритм частотою 40 Гц, що може бути пов'язаний із міжмодальним зв'язуванням сенсорної інформації; або вони дослідили, що ретикулярна активаційна система тримає людей у стані чування. Цей приклад крайній. Типові ж помилки ще непомітніші, хоч вони належать до того ж роду помилок. Що стосується свідомості, людей найбільше цікавить рефлексія, самоусвідомлення, розуміння, що людина у дзеркалі, — це «я», а також важка проблема суб'єктивного досвіду, як її описав Девід Ча́лмерс. У контексті добових ритмів прикметником «свідомий» ми описуємо ще стан чування як протилежність спанню. Але все це різні поняття, що ховаються під однією лемою. І ці засадничі явища — різні загадки науки. Стан чування ви можете пояснити без рефлексії чи суб'єктивності.

Хиба стиснення лежить в основі філософської техніки «заманюй і замінюй»: спочатку говорите про «свідомість» в одному значенні (як-от здатність думати про мислення), а потім переходите до висновків про «свідомість» в іншому (як-от суб'єктивність). Звісно, може виявитися так, що ці два явища — одне, але якщо воно справді так, для справжнього *розуміння* потрібно

спочатку розділити поняття, а *потім* після проблиску інтуїції знову об'єднати.

Розширення мапи (повторюся) — ще та *наукова* задача — частина мистецтва науки, навичка досліджувати світ. (А наукові загадки — ви не можете розв'язати наукову проблему, апелюючи до словників, так само як не можете опанувати складну навичку дослідження, кажучи: «Я можу визначати слово, як мені заманеться».) Якщо ви бачите щось заплутане із мінливими та суперечливими рисами, можна з великою ймовірністю припустити, що ваша мапа скупчує забагато інформації в одній точці — вам потрібно розділити її та виділити кілька нових категорій. Ідеться не про *визначення* того єдиного об'єкта сприйняття, але здебільшого результат *випливає* з того, що ви знаходите спосіб говорити про цей об'єкт, не використовуючи жодного уявного орієнтира.

Отже, вміння розкладати мапу на частини пов'язане з раціоналістською версією гри «Табу» та розсудливим використанням слів; адже слова часто позначають точки на нашій мапі — позначки, за якими ми класифікуємо свої твердження, — та категорії, між якими ми розподіляємо інформацію. Часто вміння уникати певного слова або вводити нові — це частина навички розширення мапи.

Категоризація має наслідки

Серед численних генетичних різновидів та мутацій, що зберігаються у вашому геномі, ви, ймовірно, знаєте лише про кілька алелів, зокрема ті, що визначають вашу групу крові: наявність чи відсутність антигенів A, B та +. Переливання крові з антигенами, відсутніми у вас, призведе до алергічної реакції. Про групи крові і те, як визначається сумісність між ними, вперше дізнався Карл Ландштáйнер (у 1930 році він отримав Нобелівську премію з медицини), і ця знахідка дала змогу безпечно переливати кров. Ба більше, якщо мати з групою A, наприклад, виношує дитину з групою A+, мати може набути алергічну реакцію на антиген +; якщо ж вона виношуватиме ще одну дитину з групою A+, дитина опиниться в небезпеці, якщо під час вагітності мати не прийматиме антиалергени. Тому перед одруженням люди перевіряють свою групу крові.

А *ще* — люди з групою A щирі і креативні, тоді як люди з групою B нестримні та життєрадісні. Люди з групою O згідливі й товариські, тоді як люди з групою AB холодні й стримані. (Може виникнути думка, що O — це відсутність A та B, тоді як AB — то A плюс B, але ні...) Усі ці передбачення пропонує нам японська теорія особистості за групою крові. Може навіть скластися враження, що група крові відіграє в Японії ту ж роль, що й знаки зодіаку на Заході — аж до гороскопів за групою крові у щоденних газетах.

Ця мода виглядає особливо дивною, позаяк група крові *ніколи не була* чимось таємничим чи то в Японії, чи то деінде. Завдяки Карлу Ландштайнеру ми взагалі знаємо про *існування* таких груп. Ані шаман, ані шанований чаклун і слова не сказав про групи крові; ні в яких стародавніх запилюжених сувоях про це теж не йшлося, навіть під спробою приховати помилку аурою старовини. Якби завтра медики заявили, що все це виявилося грандіозним обманом, такі миряни, як ми, зі своїми

неозброєними органами чуття не нашкребли б і йоти свідчення проти їхніх спростувань.

Між групами крові ніколи не було війни. Чи політичного конфлікту. Тож ці стереотипи, далебі, постали *лише* з *самого існування* позначень.

Тепер хтось неодмінно вкаже, що ця історія — про категоризацію людей. Чи станеться те саме, якщо категоризувати рослини, каміння або офісні меблі? Не пам'ятаю, чи читав про такий експеримент, що не значить, що такого ніхто не проводив. (Припускаю, головна складність для його проведення полягала б у підборі протоколу, котрий не збивав би з пантелику піддослідних, змушуючи їх думати, що якщо їм дали якесь позначення, то воно має бути якимось важливим.) Не кажу оновлюватися на основі уявних свідчень — очікую, що результат такого експерименту був би позитивним: виявили б, що сам процес позначування має силу над усіма об'єктами (принаймні в людській уяві).

Це можна побачити зі скупчення подібностей: щойно ви проводите межу довкола групи, розум відтепер намагається виявити спільні риси всередині цієї групи. На жаль, людські системи розпізнавання закономірностей функціонують настільки інтенсивно, що ми починаємо бачити ті закономірності незалежно від їх наявності: через вибіркову пам'ять слабку негативну кореляцію можна сплутати з сильною позитивною.

Це можна побачити і з нейронних алгоритмів: називати множину елементів — ніби виділяти підмережу для пошуку закономірності в ній.

Це можна побачити і з помилки стиснення: елементи, яким дали однакове ім'я, зрештою скупчаться в одному інтелектуальному контейнері, а на мапі зіллються в одну точку.

Або побачити з необмеженої людської здатності щось висмоктувати з пальця і вірити в це — ніхто ж усе одно не доведе хибність цього. Щойно ви називаєте категорію, ви можете почати вигадувати щось про неї. Найменована категорія може виходити за межі органів чуття, їй не обов'язково існувати і навіть бути зв'язною.

І це явище не обмежується Японією. Тут, на Заході, книга «Правильна дієта під вашу групу крові» (Eat Right 4 Your Type[1]) стала бестселером.

З якого боку не подивись, а окреслення межі у просторі об'єктів — не нейтральний крок. Можливо, виключно баєсівський ШІ з чіткішим дизайном зміг би проаналізувати довільний клас і не запраймитися об нього. А ось ви, людина, не маєте такого вибору. У мозку категорії динамічні: щойно ви подумали про якусь із них, вони вплинули на ваш розум. Ось іще одна причина перестати вірити, що ви можете визначити слово, як вам заманеться.

1. https://en.wikipedia.org/wiki/Blood_type_diet

Приховані конотації

У попередньому есеї ми дізналися, що в Японії місце астрології посіли групи крові: якщо, наприклад, ваша група АВ, ви, очікувано, «холодні і стримані».

Нехай ми придумаємо нове слово — «віґґін» — і будемо *називати* ним людей із зеленими очима і чорним волоссям...

> Зеленоокий чоловік із чорним волоссям зайшов у ресторан. — Ти подиви, — гукнув Денні з сусіднього столика, — бачила? У нас тут віґґін. У печінках вже ті віґґіни. Усе промишляє, де б що утнути. Його сестра Ерда зітхнула: — Денні, а ти *бачив* хоча б раз, як він щось зле утнув? — А мені й не треба, — відказав Денні і пустився по словнику. — Гляди: оно навіть в Оксфордському словнику написано: «ВІҐҐІН. 1. Людина із зеленими очима і чорним волоссям». У нього зелені очі і чорне волосся — він віґґін. Чи ти будеш сперечатися з Оксфордським словником? *За визначенням*, зеленоока чорноволоса людина — то віґґін. — Але ж ти назвав його віґґіном, — сказала Ерда. — По-хамськи казати щось про людину, коли навіть не знаєш її. У тебе ж немає свідчень, що він ллє забагато кетчупу у бургер чи, як був дитиною, стріляв у білочок з рогатки. — Але він *віґґін*, — і далі вів своє Денні. — У нього зелені очі і чорне волосся. От побачиш: щойно йому принесуть бургер, він потягнеться за кетчупом.

За посередництва слова зі спостережених ознак розум перестрибує до висновуваних. У силогізмі «Усі люди смертні, Сократ — людина, отже, Сократ смертний» до спостережених ознак входять одяг Сократа, його мовлення, вміння орудувати інструментами і людська комплекція тіла. Сократ класифікується як

«людина». Висновувана ознака — можливість отруїтися болиголовом.

Певна річ, відмінність між «спостережуваною ознакою» та «висновуваною ознакою» не така вже строга. Якщо ви чуєте, як хтось говорить, за інших рівних умов хтось може просто виглядати як людина. Якщо ви бачите тінь людини, за інших рівних умов, імовірно, вона може й говорити.

І все ж деякі властивості, як правило, частіше висновують, ніж спостерігають. Ви радше зробите висновок, що перед вами людина, а отже, вона згорить від відкритого вогню, ніж проведете цей умовивід у зворотному напрямку: побачивши, що щось горить, зробите висновок, що це людина.

Якщо зазирнути у словник за визначенням слова «людина», скоріше за все, ви натрапите на такі риси, як «інтелект», «безпір'яність» і «двоногість», помічні для швидкого візуального розрізнення людей від нелюдей. Вам не наводитимуть десятка тисяч конотацій — від вразливості до болиголова до надмірної впевненості, — які ми змогли б виснувати з категорії приналежності «бути людиною». Чому? Певно, завдання словників — дати вам зіставити позначення з групами подібностей і таким чином швидко виокремити скупчення у просторі можливих об'єктів. Чи, може, основні розрізнювальні риси, найбільш помітні, а тому й першими спадають на думку лексикографу. (Не певен, наскільки лексикографи свідомі того, що вони *насправді* роблять.)

Та зрештою виходить так, що в Оксфордському словнику Денні бачить лише перші примітні ознаки, котрі виділяють віґґінів — зелені очі та чорне волосся. Але численні дрібні *конотації*, котрі з часом засоціювалися із цим словом, як-от схильність до злочинності, кулінарні особливості та деякі необачні дитячі витівки, вже не згадуються в Оксфордському словнику.

А звідки взагалі з'явилися ті конотації? Можливо, жив-був собі віґґін із такими рисами. Чи, може, хтось взяв їх зі стелі і написав низку бестселерів («Віґґін», «Розмова з віґґінами», «Ростимо вашого малого віґґіна», «Віґґіни у ліжку»). Можливо,

навіть самі віґґіни вірять у ці стереотипи і поводяться відповідно. Щойно ви називаєте якихось людей «віґґінами», це слово починає розростатися конотаціями.

Але не забувайте про урок із «Притчі про болиголов»: якщо ми керуємося логічними визначеннями класів, то ніколи не зможемо віднести Сократа до класу «людина», поки не побачимо, як він помре. Щоразу, коли хтось дістає словник, як правило, цей хтось намагається додати *відтінки* значення, яких навіть не записано у тому словнику.

Та й, зрештою, якщо *єдине* значення слова «віґґін» зводиться до «людини із зеленими очима і чорним волоссям», чому б тоді не так і називати їх? А якщо вам цікаво, чи комусь подобається кетчуп, чому б просто не запитати: «Йому подобається кетчуп?» замість «Він віґґін?». (Згадайте про заміну символу суттю.)

Ох, а щоб говорити про *справжню* проблему, потрібне *зусилля*. Так вам би довелося на власні очі побачити, як віґґін тягнеться за кетчупом. Або, може, пошукати статистичні дані про те, скільки зеленооких чорноволосих людей полюбляють кетчуп. У будь-якому разі, якщо ви сидите у вітальні із закритими очима, у вас нічого не вийде. А ще — люди ледачі. Вони радше наводитимуть аргумент «за визначенням», особливо якщо думають, що «можуть визначати слово як заманеться».

Та *справжня* причина їхньої зацікавленості чиєїсь приналежності до «віґґінів» — конотація (відчуття, що супроводжує слово), якої немає у визначенні, яким вони *заявляють*, що ним користуються.

Уявімо, Денні каже: «Дивись: у нього зелені очі і чорне волосся. Він віґґін! Ось же у словнику написано! *Отже*, у нього чорне волосся. От тепер і сперечайся з цим!».

Звучить уже не так тріумфально. Правда ж? Якби справжня мета суперечки *визначалася* словниковим визначенням — якби суперечка *була* логічно валідною, — тоді вона *відчувалася* б порожньою: або не давала б жодної нової інформації, або ходила б тавтологічними колами.

Тому це лише спроба додати конотації, котрі *не* прописані у визначенні явно, внаслідок чого виникає відчуття, що так можна *заробити собі бал*.

Аргумент «За визначенням»

«У цієї обскубаної курки дві ноги і немає пір'я — отже, за визначенням, вона людина!»

Коли люди сперечаються про визначення, за звичкою вони починають з якогось помітного, знайомого чи принаймні широко знаного набору ознак. Опісля вони дістають словник і тицяють пальцем у ці ж ознаки, котрі відповідають словниковому визначенню. Так і приходять до висновку: «Отже, атеїзм, *за визначенням*, є релігією!».

Але часто буває так, що справа геть не у помітних, знайомих, широко знаних ознаках. Сама думка про те, що двоногість Сократа достатньо очевидна для засновку в аргументі — «Отже, Сократ, *за визначенням*, — людина!», — вказує на те, що навряд чи *на кону* та двоногість (хіба що співрозмовник відповість: «Як це Сократ двоногий? Доведи!»).

Тепер у тому, щоб законно перейти від очевидних ознак до не таких уже й очевидних, проглядається важливий сенс. Ви цілком можете побачити, що Сократ має людську комплекцію тіла, і виснувати його вразливість до болиголова. Але це ймовірнісне висновування спирається не на словникові визначення чи поширений ужиток, а на всесвіт з емпіричними скупченнями подібних об'єктів.

І побудова цих скупчень не зміниться залежно від ваших визначень. Навіть якщо у словнику ви знайдете визначення «людини», в якому йтиметься про «усіх двоногих без пір'я, крім Сократа», воно не змінить того *ступеня*, за яким Сократ подібний до решти наших двоногих без пір'я.

Коректний (з погляду сукупності подібностей) аргумент звучав би так: «Сократ має дві руки, дві ноги, ніс, язик, він вільно говорить давньогрецькою, орудує інструментами, і в кожній спостережуваній рисі він, здається, має усі основні та другорядні властивості, котрі характеризують його як *Homo sapiens*, тож я роблю припущення, що у нього людська ДНК, біохімія, а ще він

вразливий до болиголова, як і всі інші *Homo sapiens*: для них, як видно з клінічних досліджень, прийом болиголова закінчується смертю».

Нехай я відповім: «А я от бачив, як Сократ із травниками ходили по полях і, як мені здалося, збирали трави для протиотрути. Тому я *не* очікую, що, випивши болиголов, Сократ звалиться з ніг, а отже, стане винятком із загальної поведінки об'єктів у його скупченні, позаяк вони не прийняли протиотруту, а він прийняв».

Тепер не так уже й важливо, «людина» Сократ чи ні. Розмова має перейти на більш предметний рівень: ви маєте *покопатися у* деталях, завдяки яким і постала категорія «людина»: почати говорити про біохімію людини, а особливо її нейротоксичну реакцію на конііїн.

Якщо ви і далі наполягатимете: «Але ж Сократ — людина, а люди, *за визначенням*, смертні!», тоді насправді ви намагаєтеся розмити вагу ваших знань про Сократа, *крім* його людськості. Ви наполягаєте, що єдине коректне передбачення — те, яке ви зробили б, якби анічогісінько не знали про Сократа, *крім* його приналежності до людей.

Так само можна наполягати, що монетка з імовірністю 50% випаде числом або гербом, тому що вона «врівноважена», уже після того, як ви *побачили саму монетку* і що вона випала числом. Все одно що наполягати, що у Фродо десять пальців, бо у більшості хобітів десять пальців, уже після того, як ви вже *подивилися на його руки* і побачили дев'ять пальців. По-хорошому, за законами баєсівської теорії ймовірностей, так робити незаконно: ви не можете просто не обумовлювати за новими свідченнями.

Ви не можете отак просто триматися однієї категоризації і проводити розрахунки на її основі, а тим часом викидати решту того, що знаєте.

Певна річ, не кожне нове свідчення кардинально змінює ситуацію. Якщо я побачу у Сократа дев'ять пальців, це спостереження, вочевидь, не змінить моєї оцінки щодо його вразливості до

болиголова, оскільки я очікую, що втрата одного пальця не змінить його біохімію. І ця істина не зміниться *залежно від того, чи* словникове визначення якось згадує кількість пальців у людей. Законний висновок ґрунтується на скупченій структурі середовища і причиновій структурі біології — *не* на записах лексикографів чи навіть «поширеному вжитку».

Зазвичай, коли ви *правильно* робите висновок, себто *законно*, скажіть: «Алкалоїд коніїн у болиголові призводить до паралічу м'язів, що закінчується смертю удушенням». Простіше: «Люди вразливі до болиголова». Так зазвичай кажуть у *належній* суперечці.

А коли ж хтось мав би відчути потребу *посилити* аргумент підсилювальним словосполученням «за визначенням»? (Наприклад: «Люди вразливі до болиголова *за визначенням*!».) Коли висновувана ознака була поставлена під сумнів, Сократ порадився з травниками, так що мовець може відчувати потребу затягнути лещата логіки.

Тож коли ви бачите «за визначенням» у такому контексті, зазвичай це словосполучення означає: «Забудь, що ти чув, як Сократ радився з травниками, — люди, *за визначенням*, смертні!».

Люди відчувають потребу стиснути суперечку до єдиного напрямку: «Будь-яке P, за визначенням, має властивість Q!» — саме в тих випадках, коли вони бачать і воліють відкидати *додаткові аргументи*, котрі ставлять під сумнів процес висновування за замовчуванням, що ґрунтується на скупчуванні.

Та сама історія і з аргументом типу «X, *за визначенням*, — Y!». Наприклад: «Атеїсти вірять в неіснування Бога; отже, атеїсти мають віру щодо Бога, а негативна віра — теж віра; отже, атеїзм дає відповіді на теїстичні питання; отже, атеїзм, *за визначенням*, — релігія».

У вас немає потреби сказати: «Індуїзм, *за визначенням*, — релігія!», бо, ну, звісно ж, індуїзм — це релігія. Не просто релігія «за визначенням», а *власне* релігія.

Атеїзм не нагадує центральних членів скупчення «релігія», тож якби не те, що атеїзм — це релігія, *за визначенням*, ви, певно, подумали б, що атеїзм *не є* релігією. Ось тому вам доводиться розбивати опозицію ствердженням, що «Атеїзм — релігія» — істинне *за визначенням*, бо ніяким іншим чином істинним воно не є.

Ось що я хочу сказати: люди наполягають на тому, що «X, *за визначенням*, — Y!», у тих випадках, коли вони намагаються додати конотацію Y, котра прямо не прописана у визначенні, а X не виглядає як типовий член скупчення Y.

Останні тринадцять років я стежу за тим, як часто це словосполучення вживають правильно, а як часто — неправильно, хоча, боюся, не в буквальному статистичному сенсі. Але якщо видати на-гора, то використання словосполучення «*за визначенням*», за винятком математики, — найтривожніший, з мого досвіду, сигнал аргументу, зуроченого логічними хибами. Воно на ряду зі згадками «Гітлер[1]», «Бог», «абсолютно певен» і «не можеш довести, що».

Ця евристика провалу аргументу не досконала: уперше я побачив коректний ужиток за межами математики у Річарда Фейнмана, і відтоді знаходив іще. Але, певно, краще викинути «за визначенням» зі свого лексикону, а *особливо* в тих ситуаціях, коли у вас виникне спокуса вжити його курсивом чи додати після нього знак оклику. Ось ця ідея кепська *за визначенням!*

1. https://uk.wikipedia.org/wiki/Закон_Годвіна

Де ж провести межу?

Хтось підходить до вас і каже:

> «Я довго міркував над значенням слова «мистецтво» і нарешті знайшов задовільне визначення: "Мистецтво — це щось, що створюється для отримання реакції від авдиторії"».

Лишень те, що слово «мистецтво» існує, не означає, що воно ***несе якесь значення****, що витає десь у порожнечі і яке ви можете* ***відкрити*** *для себе, знайшовши правильне визначення.*

Воно так відчувається, але насправді так не є.

Питаючи, *яке визначення дати слову*, ви дивитеся на проблему не з того боку: ви шукаєте таємничу сутність того, що насправді являє собою комунікаційний сигнал.

Отже, *існує* реальний виклик, який раціоналіст цілком справедливо може прийняти, але проблема не в підборі задовільного визначення для слова. Справжню проблему можна розіграти й одному і навіть не випустити ані пари з вуст. Ця проблема — з'ясувати, котрі об'єкти подібні одне до одного — тобто котрі з них скупчуються — та іноді котрі з них мають спільну причину.

Якщо у визначення «електромагнетизму» включити блискавку, компаси, виключити з нього світло і включити месмеризм (який сьогодні ми називаємо гіпнозом), стане дещо складніше відповісти на питання «Як працює електромагнетизм?». Адже ви зліпили докупи об'єкти, котрі не належать до спільного скупчення, і виключили необхідні для формування достатньої множини. (Цей приклад історично правдоподібний: Франц Ме́смер з'явився раніше Майкла Фарадея.)

Можна було б сказати, що електромагнетизм — *не те слово*, хибно проведена межа у просторі об'єктів, що закручується петлею і звивається крізь скупчення, розріз, нездатний розсікти дійсність уздовж її природних швів.

Тому задача, *гідна раціоналіста*, — з'ясувати, де різати дійсність, щоб розсікти вздовж природних швів. Ось до чого *мали б* іти люди, коли вирушають на пошуки примарної суті слова.

Не переплутайте: усвідомити, що для опису дихання і вогню потрібне одне слово[1], — *наукова* задача. Тому не думайте звертатися до лексикографів, позаяк це не їхня робота.

Що таке «мистецтво»? Не існує ніякої суті слова, що ширяє у безодні.

Можливо, ви навіть прийдете до мене з довгим переліком того, що ви звете «мистецтвом» і «немистецтвом».

> «Маленька фуга соль мінор»: мистецтво Удар в ніс: немистецтво Літографія «Відносність» Мауріца Ешера: мистецтво Квітка: немистецтво Мова програмування Пайтон: мистецтво Плаваючий у сечі хрестик: немистецтво Цикл «Тшай» Джека Венса: мистецтво Сучасне мистецтво: немистецтво

А ви мені скажете: «Мені здається інтуїтивним провести цю межу, хоч і не знаю чому. Чи могли б ви знайти зміст поняття, що підпадає під його обсяг? Чи могли б ви сформулювати для мене *простий* опис цієї межі?».

Я відповідаю: «Гадаю, тут ідеться про захоплення ремеслом: вкладаєш працю — отримуєш диво. Для усіх цих елементів спільне те, що вони викликають схожі естетичні емоції, і те, що в їх створення люди вклали свідомі зусилля, щоб викликати саме такі емоції».

То моя відповідь допомагає, чи я просто обходжу правила раціоналістського табу? Я б сказав, що перелік людських емоцій, *естетичних* чи *неестетичних*, набагато коротший за перелік усього того, що є мистецтвом чи не є мистецтвом. Ці емоції можна побачити на знімку функціональної магнітно-резонансної

1. І полум'я, і організм споживають кисень. По суті, з погляду хімії, дихання — це дуже повільне контрольоване горіння. — Прим. пер.

томографії (фМРТ) — наголошую: вони не якісь абстрактно-духовні.

Але ж суть не в моєму визначенні мистецтва. Суть у тому, що ви легко можете розкритикувати чи то зміст, чи то обсяг мого визначення.

Ви могли б сказати: «*Не* естетична емоція спільна для всіх цих елементів. Спільний — намір пробудити *будь-яку* складну емоцію заради самого пробудження». Це судження оскаржуватиме змістовість мого визначення — мою спробу провести криву через точки даних. Ви б сказали: «Можливо, ваше рівняння й наближено відповідає цим точкам, але це не той розподіл, що їх породив».

Або ви могли б оскаржити обсяговість мого визначення: «Деякі з цих елементів пов'язані між собою — розумію, до чого ви ведете, — але Пайтону не місце у цьому списку, а сучасному мистецтву якраз місце». (Так ви показали б себе як довірливого обивателя, але можете з цим посперечатися.) Тут припускається, що дійсно існує якась базова крива, що формує цей уявний перелік подібних і неподібних елементів — що в цьому є певна закономірність, *хоча ви ще не визначили, звідки вона береться*, — але я несвідомо втратив ритм і включив деякі точки даних з іншого джерела.

Задовго до того, як вам стає *відомо*, що спільного між електрикою і магнетизмом, ви все ще можете підозрювати — на основі зовнішніх ознак, — що «месмеризм» не належить до цього переліку.

Якось вважали, що дельфінів можна було назвати «рибами». Тепер ви можете увімкнути кмітливого сперечальника і сказати: «Список: {лосось, гуппі, акули, дельфіни, форель} — всього лише список: не можна сказати, що він *хибний*. Теорією множин я можу довести, що цей список існує. Тож моє визначення *риби* — котре зводиться до простого обсягового списку — не може бути, як ви кажете, хибним».

Або ви можете перестати корчити з себе дурника і визнати, що дельфіни не належать до риб'ячого списку.

Ви придумали список елементів, що *здаються* подібними, і робите припущення, яка тому причина. Але коли ви зрештою відкриваєте для себе, що між ними *справді* спільного, ваша здогадка може виявитися хибною. Навіть ваш список може виявитися хибним.

Ви не можете ховатися за зручним щитом «правильно за визначенням». Що обсягові, що змістові визначення можуть хибувати — розрізати дійсність не за її швами.

Категоризування — це спекулятивні зусилля, при яких ви можете припускатися помилок. Тому мудро могти визнати (чисто теоретично), що ваші здогадки про визначення можуть бути «помилковими».[2]

2. На поняття «межі» добре лягає поняття *межі рішень* з машинного навчання. У просторі ознак вона розділяє області, яким класифікатор приписує різні класи. У цьому сенсі завдання раціоналіста можна уявити як пошук такої межі, яка не просто правильно класифікує вже відомі приклади, а й узагальнюється на нові спостереження та відображає структуру даних: їхні спільні причини, закономірності й відмінності у прогнозах. Ця аналогія водночас показує, чому «правильну» межу не обов'язково можна просто виявити в самих об'єктах. Різні межі можуть однаково добре класифікувати наявні дані, але по-різному поводитися на нових. Надто складна межа може просто запам'ятати окремі приклади (перенавчитися (https://uk.wikipedia.org/wiki/Перенавчання)), тоді як надто проста може не відобразити важливу структуру даних. Тому корисність категорії визначається не лише тим, наскільки добре вона описує відомі випадки, а й тим, наскільки добре вона узагальнюється, допомагає робити прогнози та відповідає меті, заради якої ми її використовуємо. — Прим. пер.

Ентропія і короткі коди

Нехай у вас є система X, що рівноймовірно може перебувати у восьми можливих станах:

$$\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8.\}$$

У фізиці, математиці та навіть біології надзвичайно поширена величина, що зветься *ентропією*. І ентропія системи X дорівнює трьом бітам. Тобто, щоб знайти значення X, у середньому нам потрібно поставити три запитання із відповіддю «так або ні». Наприклад, хтось міг би підказати нам значення X, використовуючи цей код:

X_1: 001	X_2: 010	X_3: 011	X_4: 100
X_5: 101	X_6: 110	X_7: 111	X_8: 000

Якби я запитав: «Перший символ 1?» — і почув «так», тоді запитав би: «А другий символ 1?» — і почув «ні», тоді запитав би: «Третій символ 1?» і почув «ні», я б дізнався, що система X перебуває у четвертому стані.

Тепер скажімо, що система Y має чотири можливих стани з наступними ймовірностями:

Y_1: 1/2(50%) Y_2: 1/4(25%) Y_3: 1/8(12,5%) Y_4: 1/8(12,5%)

Тоді ентропія Y становила б один цілий сімдесят п'ять сотих біта, тобто дізнатися її значення коштувало б 1,75 запитання.

Що це взагалі означає — поставити одне і три чверті питання? Уявіть, що ми позначаємо стани Y наступним кодом:

Y_1: 1	Y_2: 01	Y_3: 001	Y_4: 000

Спершу ви запитаєте: «Перший символ 1?». Відповідь «так» — результат: Y у першому стані. Так буває у 50%

випадків — потрібне одне запитання «так/ні», щоб дізнатися його стан.

А от тепер нехай відповідь буде «ні». Тоді ви питаєте: «Другий символ 1?». «Так» — результат: система у другому стані. Ймовірність перебування у цьому стані складає 1/4, і щоразу, коли він настає, ми отримуємо цю інформацію з двох питань із відповіддю «так/ні», тож у 25% випадків для цього знання знадобиться два питання.

Якщо двічі підряд відповідь «ні» — питаєте: «Третій символ 1?». «Так» — Y у третьому стані; «ні» — значить у четвертому. В одному випадку з восьми Y перебуває у третьому стані, і тоді ця інформація коштує три запитання; в одному випадку з восьми Y — у четвертому стані, що коштує трьох запитань.

$$(1/2 \times 1) + (1/4 \times 2) + (1/8 \times 3) + (1/8 \times 3)$$
$$= 0{,}5 + 0{,}5 + 0{,}375 + 0{,}375 = 1{,}75.$$

Загальна формула для ентропії системи S — це сума по всіх S_i по $-p(S_i) \times log_2(p(S_i))$.

Наприклад, логарифм 1/8 з основою 2 дорівнює -3. Тож $-(1/8 \times (-3)) = 0{,}375$ — це вклад стану S_4 до загальної ентропії: в одному з восьми випадків нам доведеться поставити три питання.

Не завжди можна написати досконалий код для системи, але якщо вам треба одним повідомленням передати комусь стан довільної кількості копій S, ви можете довільно наблизитися до бездоганного коду. (Загугліть «арифметичне кодування» для простого методу.)

Тут ви можете запитати: «Чому для Y_4 замість 000 не написати 10? Чи не швидше передати повідомлення отак?».

Але якщо для Y_4 ви напишете 10, тоді, коли хтось відповідає «так» на питання «Перший символ 1?», ви не дізнаєтеся, чи система у стані Y_1 (1) чи Y_4 (10). Якщо ви так зміните код, розвалиться ціла система, оскільки: почуєте «1001» — не дізнаєтеся, чи він означає «за Y_4 іде Y_2» чи «за Y_1 іде Y_3».

Мораль: *короткі слова — це обмежений ресурс*.

Ключ до написання хорошого коду — що передає повідомлення якомога лаконічніше, — приберегти короткі слова для об'єктів, про які доведеться говорити часто, і довші слова — для тих, про які не доведеться говорити так часто.

Якщо довести це мистецтво до його межі, довжина повідомлення, потрібна для опису чогось, точно чи майже точно відповідатиме її ймовірності. Так з'являється мінімальна довжина опису, чи мінімальна довжина повідомлення, — формалізація леза Оккама.

А відтак навіть *позначення* слів не повністю довільні. Звуки, які ми надаємо нашим поняттям, можуть бути доречнішими чи недоладнішими. Навіть якщо не враховувати поширеного вжитку!

Я пишу усе це, щоб сказати: ідея «Ви можете робити *X* як заманеться» — значна перешкода на шляху до розуміння того, як мудро робити *X*. «У нас вільна країна — я маю право на власну думку» — завада у мистецтві знаходженні істини. «Я можу визначати слово як мені заманеться» — завада у мистецтві розкроювання дійсності по її природних швах. І навіть буцімто притомне «Позначення, які ми чіпляємо до слів, довільні» заважає усвідомити компактність. До речі, сюди ж іде й просодія[1]: якось Толкін зауважив[2], завдяки чому словосполучення cellar door таке милозвучне. Саме така усвідомленість потрібна, що користуватися мовою на рівні Толкіна.

Довжина слів теж грає нетривіальну роль у когнітивному аспекті мовознавства.

От візьмімо словосполучення «крісло з підставкою» та слова «стілець» і «меблі». Крісло з підставкою стосується конкретнішої категорії, ніж просто «стілець», тоді як «меблі» — то

1. Сукупність інструментів, якими озвучують мову (інтонація, гучність, висота звуку, темп, наголоси та павзи). — Прим. пер.

2. У своїй лекції «Англійська та валлійська мови» Толкін згадав (https://tolkiengateway.net/wiki/Cellar_door) словосполучення cellar door як приклад особливо гарного поєднання англійських слів. — Прим. пер.

ще загальніше за «стілець». Проте значна більшість стільців мають спільне призначення: щоб сісти на них, потрібно виконати однакові рухи, і сідають на них з однаковою метою (щоб змістити тягар з ваших ніг, поки ви їсте, читаєте, пишете чи відпочиваєте). Крісло з підставкою не сильно відрізняється у цьому плані. А от «меблі» охоплюють такі об'єкти, як ліжка і столи, у яких різне призначення, і, на відміну від стільців, задіюють різні рухи.

У термінології когнітивної психології «стілець» — *категорія базового рівня*.

Як правило, люди говорять і, ймовірно, думають на базовому рівні категоризації: вони окреслюють межу довкола поняття «стілець», а не конкретнішого поняття «крісло з підставкою» чи загальнішого поняття «меблі». Людина ймовірніше скаже: «Сідайте собі на стілець», а не «Сідайте собі на крісло з підставкою» чи «Сідайте собі на меблі».

Та й не назвати збігом те, що у слові «стілець» менше складів, аніж у словосполученні «крісло з підставкою» чи «меблі»[3]. Загалом, як правило, категорії базового рівня мають короткі назви, а короткі іменники зазвичай позначають базові категорії. Ясно, що правило не залізобетонне, але певна тенденція, безперечно, простежується: частотний слововжиток супроводжується короткими словами і навпаки.

Як казав Дуґлас Гофстедтер, не без причини в англійській мові під артиклем the мають на увазі саме визначений артикль, а під «антидісестаблішментаризмом» мають антидісестаблішментаризм, а не навпаки.

3. Приклад ілюстративний. В англійській мові слово *furniture* /ˈfɜː.nɪ.tʃɚ/ має три склади, на відміну від українського «меблі», де лише два склади — [мéбл'і]. Також в українській мові слово [стілéц'] має два склади, як і «меблі», на відміну від англійських /tʃer/ (1) та /ˈfɜː.nɪ.tʃɚ/ (3). — Прим. пер.

Взаємна інформація та щільність у просторі об'єктів

Нехай у нас є система X з вісьмома рівноймовірними станами (відносно поточного стану ваших знань), а система Y може існувати у чотирьох рівноймовірних станах.

Як уже проговорював у попередньому есеї, ентропія X вимірюється трьома бітами: щоб дізнатися її поточний стан, знадобиться три питання з відповідями «так або ні». Ентропія ж Y вимірюється двома бітами: щоб дізнатися її поточний стан, відповідно, потрібно два питання. Здається очевидним, бо $2^3 = 8$, а $2^2 = 4$, тож три питання дозволяють відокремити вісім можливостей, а два питання — чотири. Та не забувайте: якби можливості були не рівноймовірними, ми змогли б написати розумніший код для визначення стану Y за допомогою у середньому, наприклад, 1,75 питання. Хоча у цьому випадку *імовірнісна маса X рівномірно розподілена* по всіх її можливих станах, як і в Y, тож ні про які розумні коди говорити не доводиться.

Якою тоді буде ентропія об'єднаної системи (X, Y)?

Може виникнути спокуса відповісти: «Для X потрібно три питання, для Y — два, тож сумарно знадобиться п'ять питань для з'ясування стану X та Y».

А що як ті дві змінні сплутані, внаслідок чого знання про стан Y може щось нам сказати і про стан X?

Зокрема, скажімо, що X та Y або обидва парні, або обидва непарні.

Якщо ми отримаємо трьохбітне повідомлення (поставимо три питання) і дізнаємося, що X у п'ятому стані, ми зрозуміємо, що Y у першому чи третьому стані, але не в другому або четвертому. Тоді лишається ще одне додаткове питання: «Y у третьому стані?» — відповідь «ні» дає нам знати про стан цілої системи (X, Y): $X = X_5$, $Y = Y_1$. Цю інформацію ми дістали всього за чотири питання.

І навпаки — якщо з двох питань ми дізнаємося, що Y у четвертому стані, знадобиться тільки два додаткових питання про те, чи X у другому, четвертому, шостому чи восьмому стані. Знову ж — чотири питання на те, щоб дізнатися стан об'єднаної системи.

Взаємною інформацією двох змінних зветься різниця між ентропією об'єднаної системи та ентропією незалежних систем: $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

Отут і заховався один біт взаємної інформації між нашими двома системами: знання про X дає нам біт інформації про Y (обрізає простір можливостей з чотирьох до двох — зменшує об'єм удвічі), а знання про Y дає біт інформації про X (обрізає простір можливостей з восьми до чотирьох).

А як щодо ситуації, коли імовірнісна маса розподілена нерівномірно? У попередньому есеї, наприклад, ми розглядали випадок, де для чотирьох станів Y були визначені ймовірності 1/2, 1/4, 1/8, 1/8. Візьмімо їх для нашого імовірнісного розподілу по Y, розглянемо їх незалежно: якщо побачили Y і тільки Y, саме це ми й очікували побачити. Тепер припустімо, що змінна Z має два стани: перший і другий із відповідними ймовірностями 3/8 та 5/8.

Тоді взаємна інформація між Y та Z нульова, якщо і тільки якщо спільний розподіл Y та Z виглядає ось так:

$Z_1 Y_1$: 3/16	$Z_1 Y_2$: 3/32	$Z_1 Y_3$: 3/64	$Z_1 Y_3$: 3/64
$Z_2 Y_1$: 5/16	$Z_2 Y_2$: 5/32	$Z_2 Y_3$: 5/64	$Z_2 Y_3$: 5/64

Цей розподіл підкоряється закону:

$$p(Y,Z) = P(Y)P(Z)$$

Наприклад: $P(Z_1 Y_2) = P(Z_1)P(Y_2) = 3/8 \times 1/4 = 3/32$.

Зверніть увагу, що ми можемо встановити відособлені (незалежні) ймовірності Y та Z, просто подивившись на спільний розподіл:

$P(Y_1)$ = загальна ймовірність усіх різних способів настання Y_1 $= P(Z_1 Y_1) + P(Z_2 Y_1) = 3/16 + 5/16 = 1/2$.

Отже, лише проаналізувавши спільний розподіл, ми можемо встановити, чи відособлені змінні Y та Z незалежні — тобто чи розкладається спільний розподіл на добуток відособлених розподілів, чи для всіх значень Y та Z виконується $P(Y, Z) = P(Y)P(Z)$.

І останнє важливе, оскільки за теоремою Баєса:

$$P(Y_i, Z_j) = P(Y_i)P(Z_j) \qquad P(Y_i, Z_j)/P(Z_j) = P(Y_i)$$
$$P(Y_i|Z_j) = P(Y_i)$$

Українською: «Як тільки дізнаєтеся про Z_j, ваше уявлення про Y_i *не зміниться*».

Тож розкладання розподілу на множники — $P(Y, Z) = P(Y)P(Z)$ — рівнозначне твердженню «Інформація про Y не дасть вам жодних відомостей про Z, і навпаки».

З цього ви можете правильно запідозрити, що між Y та Z немає ніякої взаємної інформації. А де немає взаємної інформації, там зась шукати баєсівського свідчення, і навпаки.

Нехай у вищенаведеному розподілі YZ ми розглядали кожну можливу комбінацію Y та Z як незалежну подію — тобто розподіл YZ мав загалом вісім варіантів із зазначеними ймовірностями, — а потім так само порахували ентропію розподілу YZ, як і будь-якого іншого розподілу:

$$3/16\log_2(3/16) + 3/32\log_2(3/32) + 3/64\log_2(3/64) + \ldots + 5/64\log_2(5/64)$$

.

Якби ви рахували ентропію кожної системи окремо і додали їх, то, зрештою, отримали б той самий результат. Тож, якщо між цими двома змінними немає взаємної інформації, то й наша невизначеність щодо об'єднаної системи не менша від нашої невизначеності щодо цих двох систем, розглянутих окремо. (Не наводжу розрахунків, але ви цілком можете зробити їх самі; не наводжу і доведення того, що ці міркування істинні загалом, але

можете загуглити «інформаційна ентропія Шеннона» та «взаємна інформація».)

А що як спільний розподіл не розкладається на множники? Наприклад:

Z_1Y_1: 12/64	Z_1Y_2: 8/64	Z_1Y_3: 1/64	Z_1Y_4: 3/64
Z_2Y_1: 20/64	Z_2Y_2: 8/64	Z_2Y_3: 7/64	Z_2Y_4: 5/64

Якщо скласти спільні ймовірності, щоб отримати відособлені, ви побачите: $P(Y_1) = 1/2$, $P(Z_1) = 3/8$ і так далі — відособлені ймовірності ті самі, що й раніше.

Проте спільні ймовірності не завжди дорівнюють добутку відособлених імовірностей. Наприклад, імовірність $P(Z_1Y_2)$ дорівнює 8/64, де $P(Z_1)P(Y_2)$ складає $3/8 \times 1/4 = 6/64$. Тобто ймовірність того, що вам одночасно трапиться Z_1Y_2, вища за очікувану ймовірність незалежного спостереження Z_1 чи Y_2.

А звідси, своєю чергою, виливається у

$$P(Z_1Y_2) > P(Z_1)P(Y_2) \qquad P(Z_1Y_2)/P(Y_2) > P(Z_1)$$
$$P(Z_1|Y_2) > P(Z_1)$$

З того, що ймовірність $P(Z_1Y_2)$ «незвично висока» — вища, ніж передбачають відособлені ймовірності за замовчуванням, — слідує те, що спостереження Y_2 є свідченням, котре збільшує ймовірність Z_1. За симетричним міркуванням, спостереження Z_1 має підкріплювати ймовірність Y_2.

Оскільки бодай якісь значення Y дають нам інформацію про Z (і навпаки), між цими двома змінними мусить існувати взаємна інформація. Тож ви дізнаєтеся (я впевнений, хоч і не перевіряв), що обчислення ентропії YZ дає меншу сумарну невизначеність, аніж сума незалежних ентропій Y та Z. $H(Y, Z) = H(Y) + H(Z) - I(Y; Z)$, причому усі величини обов'язково невід'ємні.

(Тут я відхилюся від теми і зауважу, що симетрія виразу для взаємної інформації вказує на те, що в середньому Y

> *має* надавати нам стільки ж інформації про Z, скільки Z дає про Y. Лишаю читачеві завдання узгодити цю тезу із тим, чого його навчали на уроках логіки: якщо всі во́рони чорні, то можливість робити висновок «Ворон(x) → Чорний(x)» не означає, що так само можна виснувати «Чорний(x) → Ворон(x)». Наскільки ж симетричні потоки ймовірностей у баєсіанстві відрізняються від різких перепадів логіки, навіть якщо останнє — лише збіднений випадок першого.)

«Але ж, — все-таки запитаєте ви, — який стосунок усе це має до належного слововжитку?».

В есеях «Порожні позначення» та «Замініть символ суттю» ми ознайомилися з техніками заміни слова його визначенням. Наприклад:

> Усі [смертний, ~пір'я, двоногий] смертні. Сократ — [смертний, ~пір'я, двоногий]. Отже, Сократ смертний.

Чому ж тоді вам так хочеться мати слово для позначення «людини»? Чому б не сказати «Сократ — смертне двоноге без пір'я»?

Тому що для об'єктів, з якими ми часто маємо справу, зручно мати короткі слова. Якщо ваш код для опису одиничних властивостей вже ефективний, тоді вам не буде ніякої користі з конкретного слова для сполучення цих властивостей — наприклад, «людина» для «смертного двоногого без пір'я», — хіба що об'єкти, котрі *і* смертні, *і* двоногі, *і* без пір'я, зустрічаються *частіше*, ніж за очікуваною відособленою ймовірністю.

В ефективному коді довжина слова відповідає ймовірності: а отже, код для $Z_1 Y_2$ буде такої ж довжини, як і код для Z_1 плюс код для Y_2, хіба що $P(Z_1 Y_2) > P(Z_1)P(Y_2)$ і в такому разі код слова може бути коротшим за коди його складників.

Своєю чергою, це точно відповідає тому випадку, коли ми можемо виснувати деякі властивості об'єкта зі спостереження за його іншими властивостями. Ймовірність того, що двоногі

об'єкти без пір'я також будуть смертними, має бути вищою, ніж за замовчуванням.

Ясно, що слово «людина» описує безліч властивостей: коли ви бачите людиноподібну сутність, яка говорить і носить одяг, ви можете зробити висновки про цілу низку її біохімічних, анатомічних та мисленнєвих особливостей. Щоб замінити слово «людина» повним описом усіх наших знань про людей, знадобиться надмірно багато часу на перераховування усього. Але це так *тільки* тому, що балакуче двоноге без пір'я куди ймовірніше, ніж зазвичай, може отруїтися болиголовом, мати широкі нігті чи поводитися надто самовпевнено.

Окреме слово для об'єкта, а не просто перелік його властивостей, — це більш лаконічний код саме у тих випадках, коли ми можемо вивести деякі з цих властивостей з інших. (Певно, виняток становитимуть дуже примітивні слова типу «червоний», якими ми передаємо повністю нестиснений опис нашого чуттєвого досвіду. Але коли ви стикаєтеся із жучком чи навіть камінцем — маєте справу з непростими наборами властивостей, далеко не примітивними.)

Тож слово «віґґін» на позначення зеленооких чорноволосих людей виправданіше за проговорювання «зеленоока чорноволоса людина» саме у тих випадках, коли

1. зеленоокі зустрічаються в середньому частіше мають чорне волосся (і навпаки), себто ми можемо з певною ймовірністю зробити висновок про наявність зелених очей на підставі чорного волосся (або навпаки), *чи*
2. віґґіни поділяють інші властивості, котрі можна виснувати з ймовірністю, вищою за базову. У цьому випадку ми маємо окремо побачити зелені очі і чорне волосся; після незалежного спостереження цих двох ознак ми можемо з певною ймовірністю вивести наявність інших ознак (гастрономічну любов до кетчупу, наприклад).

Можна навіть вважати, що акт визначення слова — це обіцянка такого змісту. За максимами Ґрайса[1], казати комусь: «Я визначаю слово "вігґін" як людину із зеленими очима і чорним волоссям» — все одно що стверджувати, що слово «вігґін» допоможе прийти до якихось висновків чи скоротити ваші повідомлення.

Якщо ймовірність побачити зеленоокість та чорноволосість разом не вища за базову, і жодна інша риса не трапляється частіше, ніж зазвичай, разом із ними, тоді слово «вігґін» — брехня: воно стверджує, що деяких людей варто виокремити в групу, — а насправді не варто.

У такому випадку це слово не допомагає лаконічніше описувати дійсність — визначення йому дав не той, хто відправляє найкоротше повідомлення, — і не відіграє ніякої ролі у найпростішому поясненні. Так само «вігґін» ніяк не допоможе вам у баєсівському висновуванні. Навіть якщо не називати його брехнею, точно можна охрестити помилкою.

Адже розкроювати дійсність по її природних швах — означає окреслювати його кордони довкола скупчень із надзвичайно високою щільністю ймовірностей у просторі об'єктів.

1. У прагматиці правила ефективної комунікації, до яких входять чотири максими (https://uk.wikipedia.org/wiki/Пол_Герберт_Грайс#Філософська_праця): максима кількості, максима якості, максима релевантності та максима способу. — Прим. пер

Надекспоненційний простір понять та прості слова

Можна подумати, що простір об'єктів — надзвичайно великий простір. Набагато більший за дійсність: оскільки дійсність вміщує тільки те, що існує, простір об'єктів уміщує все, що *могло б* існувати.

Власне, спосіб, у який я «дав визначення» простору об'єктів — надав йому виміри для кожної можливої ознаки включно з такими взаємопов'язаними властивостями, як щільність, об'єм та маса, — може нечітко описувати його, тому у нього немає нічого такого, що можна було б назвати розміром. Але *так чи інакше* головне — могти уявити собі простір об'єктів. Певна річ, нікому не під силу *по-справжньому* зрозуміти зграю горобців, якщо ви бачите лише хмару істот, що махають крильцями і цвірінькають, а не скупчення точок у просторі об'єктів.

Та хай би яким широчезним не був простір об'єктів, він і на йоту не зрівняється із розмірами простору понять.

У машинному навчанні «поняття» означає правило, котре включає чи виключає приклади. Бачите 2:+, 3:-, 14:+, 23:-, 8:+, 9:- — можете здогадатися, що поняття зводиться до «парних чисел». Існує розлогий корпус літератури (чого можна було очікувати) про те, як виводити поняття з даних... на основі випадкових чи відібраних прикладів... З урахуванням можливих помилок у класифікації... і, що найважливіше, з урахуванням різних просторів можливих правил.

Наприклад, хочемо вивести поняття «підхожі дні для гри у теніс». Можливі властивості Днів такі:

Небо: {Сонячне, Хмарне, Дощове} ТемператураПовітря: {Тепло, Холодно} Вологість: {Нормальна, Висока} Вітер: {Сильний, Слабкий}

Потім нам дають наступні дані, де + позначає позитивний приклад поняття, а - — негативну класифікацію:

> + Небо: Сонячне; ТемператураПовітря: Тепло; Вологість: Висока; Вітер: Сильний. - Небо: Дощове; ТемператураПовітря: Холодно; Вологість: Висока; Вітер: Сильний. + Небо: Сонячне; ТемператураПовітря: Тепло; Вологість: Висока; Вітер: Слабкий.

Який висновок із цього має вивести алгоритм?

Програма машинного навчання може відобразити *одне* поняття, що підходить під ці дані:

> Небо: ?; ТемператураПовітря: Тепло; Вологість: Висока; Вітер: ?

У такому форматі, щоб з'ясувати, чи приймає поняття певний приклад чи відхиляє його, ми порівнюємо елемент за елементом: символ «?» приймає будь-що, проте конкретне значення приймає лише це конкретне значення.

Відтак, поняття вище приймає лише Дні з ТемператураПовітря=Тепло та Вологість=Висока, але Небо і Вітер можуть набувати будь-якого значення. Таке рішення відповідає як негативним, так і позитивним класифікаціям у наразі доступних даних, однак це поняття не *єдине*, що відповідає їм.

Крім того, ми можемо полегшити представлення вищезгаданого поняття до {?, Тепло, Висока, ?}.

Якщо не вдаватися у подробиці, класичний алгоритм виглядає десь так:

- Утримуйте набір найзагальніших гіпотез, що узгоджуються з даними, — тих, що позитивно класифікують якомога більше прикладів і при цьому все ще відповідають свідченням.
- Утримуйте ще один набір найконкретніших гіпотез, що узгоджуються з даними, — тих, що негативно класифікують якомога більше прикладів і при цьому все ще відповідають свідченням.

- Щоразу, коли бачимо новий негативний приклад, мінімальною мірою ми підкріплюємо всі найзагальніші гіпотези, так що новий набір знову якомога загальніший і все ще узгоджується зі свідченнями.
- Щоразу, коли ми бачимо новий позитивний приклад, мінімальною мірою ми послаблюємо найконкретніші гіпотези, так що новий набір знову якомога конкретніший і все ще узгоджується зі свідченнями.
- Продовжуємо, допоки залишиться одна-єдина гіпотеза. *Якщо* цільове поняття взагалі входило у наш простір гіпотез, воно й буде тією відповіддю.

У прикладі вище до набору найбільш загальних гіпотез входили б {?, Тепло, ?, ?} та {Сонячне, ?, ?, ?}, а до набору найбільш конкретних — єдиний член {Сонячне, Тепло, Висока, ?}.

Будь-яке інше поняття, що, як на вас, узгоджується з даними, буде строго більш конкретним, ніж одна з найзагальніших гіпотез, і строго більш загальним за найконкретнішу гіпотезу.

(Більше раджу вичитати з книги Тома Мітчелла «Машинне навчання» (Machine Learning), з якої я адаптував цей приклад.)

Певно, ви помітили, що формат вище *не може* відобразити всі можливі поняття. Наприклад: «Грати в теніс, коли небо сонячне *або* повітря тепле». Воно узгоджується з даними, але у цьому форматі представлення понять немає четвірки значень, що описували б це правило.

Очевидно, що наша програма штучного навчання не надто загальна. Чому б не дозволити їй відображати *всі можливі* поняття, щоб вона могла навчатися якомога гнучкіше?

Дні складаються з цих чотирьох змінних: одна змінна — з трьох значень і три — з двох. Отже, ми можемо стикнутися з $3 \times 2 \times 2 \times 2 = 24$ можливими Днями.

Запропонований формат представлення Понять дозволяє нам задавати будь-яке значення для змінної або взагалі не задавати їй ніякого значення. Отже, у такому представленні налічується $4 \times 3 \times 3 \times 3 = 108$ понять. Щоб алгоритм «від

найзагальнішого до найконкретнішого» працював, нам потрібно почати з найбільш конкретної гіпотези «жоден приклад не класифікується як позитивний». З нею всього буде 109 понять.

Чи не виглядає підозрілим те, що можливих понять більше за можливі Дні? Точно ні: зрештою, поняття можна розглядати як *сукупність* Днів. Поняття можна розглядати як множину днів, яку воно класифікує як позитивні або, ізоморфно, як множину днів, які воно класифікує як негативні.

Отже, простір *усіх можливих* понять, що класифікують Дні, — це множина усіх можливих множин Днів, розмір яких — $2^{24} = 16\,777\,216$.

Цей цілісний простір охоплює усі поняття, які ми вже розглянули. Він також включає поняття типу «Класифікувати приклади {Сонячне, Тепло, Висока, Сильний} та {Сонячне, Тепло, Висока, Слабкий} як позитивні і виключати все інше» або «Класифікувати тільки приклад {Дощове, Холодно, Висока, Сильний} як негативний і включати все інше». Він вміщує поняття, які не мають стислого представлення, — лише простий перелік усього дозволеного і недозволеного.

У цьому й проблема побудови «повністю загальної» програми індуктивного навчання: доки не побачать кожен можливий приклад у просторі прикладів, не можуть засвоїти поняття.

Додамо більше ознак до Днів (температура води чи прогноз на завтра) — число можливих днів зростатиме експоненційно за кількістю ознак. Але це не проблема для нашого обмеженого простору понять, оскільки за допомогою логарифмічного числа прикладів можна звузити широкий простір.

Додамо ознаку «Вода: {Тепла, Холодна}» до днів — отримаємо 48 можливих днів та 325 можливих понять. Нехай кожен День, що ми його бачимо, зазвичай класифікується як позитивний близько половиною з тих понять, що наразі видаються правдоподібними, а інша половина класифікується як негативна. Потім, коли ми дізнаємося дійсну класифікацію прикладу, вона скоротить простір сумісних понять удвічі. Тож може

знадобитися лише дев'ять прикладів ($2^9 = 512$), щоб звузити 325 можливих понять до одного.

Якби у Днів було сорок двійкових ознак, для звуження кола можливих понять до одного все ще знадобився б доступний обсяг даних. Шістдесят чотири приклади, якщо половину з решти понять кожен класифікувати як позитивний. Звісно, я *виходжу* з того, що *дійсне* правило взагалі можна відобразити!

Хочете розглянути усі можливості — удачі вам (що я можу сказати?). Простір *усіх можливих* понять розширюється *над*експоненційно пропорційно до кількості ознак.

Поки ви говорите про дані з сорока двійковими ознаками, число можливих прикладів сягає вже трильйона, а число можливих *понять* — двох у трильйонному степені. Щоб звузити цей *над*експоненційний простір понять, вам доведеться проглянути трильйон прикладів, перш ніж зможете визначити, що належить до нього, а що ні. Проглядати кожнісінький можливий приклад.

І це, зауважте, при сорока двійкових ознаках. Сорок бітів, або п'ять байтів, — на те, щоб визначити «Так» чи «Ні». Сорок бітів передбачають 2^{40} можливих прикладів і $2^{2^{40}}$ можливих понять, що класифікують приклади як позитивні чи негативні.

Отже, у дійсному світі, де опис об'єктів вимагає понад п'ять байтів, немає трильйона прикладів; у тренувальних даних присутній шум; ми розглядаємо лише *надзвичайно системні* поняття. Адже людський розум — чи весь видимий усесвіт — і близько не такий величезний, щоб розглянути усю решту гіпотез.

З такого погляду, навчання не просто *спирається на* індуктивний біас — воно *майже повністю* складається з нього: якщо порівняти число понять, виключених *апріорі*, із тими, що відсіюються лише на підставі свідчень.

І тут ви знову запитаєте: «А який стосунок усе це має до належного слововживання?».

У цьому й причина того, що слова мають як зміст, так і обсяг.

У попередньому есеї я прийшов до такого висновку:

> «Розкроювати дійсність по її природних швах — означає окреслювати її кордони довкола скупчень із надзвичайно високою щільністю ймовірностей у просторі об'єктів».

Я навмисно не вказав у цьому (дещо відредагованому) твердженні одне важливе уточнення, оскільки до сьогодні не міг його пояснити. Краще твердження звучало б так:

> «Розкроювати дійсність по її природних швах — означає окреслювати *прості* кордони довкола скупчень із надзвичайно високою щільністю ймовірностей у просторі об'єктів».

Інакше ви б просто зловмисно маніпулювали межами простору об'єктів. Ви б створили вкрай дивні несуміжні межі, що об'єднують спостережувані приклади, які неможливо було б описати в повідомленні, коротшому за самі спостереження, і сказали б: «Ось що я побачив раніше і те, чого я очікую побачити, буде більше у майбутньому».

У дійсному світі нічого з того, що вище рівня молекул, не повторюється *один в один*. Сократ зовні багато в чому схожий на усіх інших людей, вразливих до болиголова, але він *точно* не такий самий, як вони. Тому ваша здогадка, що Сократ — «людина», ґрунтується на окресленні простих кордонів довкола людського скупчення у просторі об'єктів. А не на «об'єкти, що повністю відповідають [п'ятимегабайтне уточнення форми 1] з [купа інших ознак] *чи* точнісінько як [п'ятимегабайтне уточнення форми 2] та [купа інших ознак], ... є людьми».

Якщо ж ви не проводите *просту* межу довкола ваших досвідів, то не зможете робити з них висновки. Тому й намагаєтеся описати «мистецтво» такими змістовими визначеннями, як «те, що призначене викликати яку-небудь складну емоцію заради неї самої», а не просто показуєте на довгий перелік об'єктів, що є мистецтвом і ним не є.

По суті, вищенаведене твердження про те, «як розкроювати дійсність по її природних швах», трохи нагадує задачу про курку та яйце: поки не зробите хоча б маленький надріз, не побачите *щільність* дійсних спостережень. А розподіл імовірностей з'являється вже з окреслення меж — не навпаки: якби у вас уже *був* розподіл імовірностей і все необхідне для висновування, чи забивали б ви собі голову окресленням меж?

І тут з'являється ще одна — так, ще одна — причина ставитися з підозрою до заяви «ви можете визначати слово як завгодно». Якщо подивитися на надекспоненційні розміри простору понять, стає ясно, що виокремити одного поняття — уже акт неабиякої сміливості не лише для нас, а й для будь-якого розуму з обмеженою обчислювальною потужністю.

Вводити в обіг слово «вігґін» (витлумачене як «чорноволоса зеленоока людина») без якої-небудь причини навмисне підносити *це конкретне поняття* до рівня, де воно потребує виправдання. Все одно що детектив скаже: «Що ж, так чи інакше у мене немає ніяковісінької зачіпки, хто міг вбити тих сиріт... навіть інтуїція мовчить... а чи брали ми до уваги Джона К. Віффлгайма з Норкл-роуд 1234 як потенційного підозрюваного?».

Умовна незалежність та наївний баєсів метод

У попередньому есеї я писав про взаємну інформацію між X та Y — $I(X;Y)$ — тобто різницю між ентропією розподілу спільної ймовірності — $H(X,Y)$ — та ентропіями відособлених розподілів — $H(X) + H(Y)$.

Навів приклад змінної X із вісьмома рівноймовірними станами 1..8, якщо ми ще не отримали яких-небудь свідчень, та змінної Y із чотирма рівноймовірними станами 1..4, якщо ми ще не отримали яких-небудь свідчень. Потім, порахувавши відособлені ентропії $H(X)$ та $H(Y)$, ми з'ясовуємо, що X має три біти ентропії, а Y — два.

Однак, ми знаємо і те, що або X та Y парні, або обидва непарні. Ось і все, що ми знаємо про відношення між ними. Отже, для спільного розподілу (X,Y) існує тільки шістнадцять рівноймовірних можливих станів при спільній ентропії чотири біти. У порівнянні з п'ятьма бітами ентропії, якби X та Y були незалежними, цей недолік ентропії становить один біт. Цей недолік і є взаємною інформацією: через X ми можемо дізнатися щось про Y чи навпаки. Тож наша непевність стосовно однієї змінної після інформації про іншу вже трохи менша.

Додамо ще й третю змінну Z. Z має два стани — парний і непарний, — що повністю корелюють з парністю та непарністю (X,Y). По суті, ми можемо припустити, що Z тотожне питанню «X та Y парні чи непарні?».

Якщо у нас немає ніяких свідчень про X та Y, сама Z обов'язково має один біт ентропії наданої інформації. Між Z та X маємо один біт взаємної інформації і між Z та Y — один біт. Як зазначалося раніше, між X та Y — один біт взаємної інформації. Тож, яка ентропія всієї системи (X,Y,Z)? Можна наївно очікувати, що

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) - I(X;Z) - I(Z;Y) - I(X;Y)$$

,

але виявляється, що справа не в цьому.

Об'єднана система (X, Y, Z) має лише шістнадцять можливих станів — оскільки Z тотожне питанню «X та Y парні чи непарні?» — тож $H(X, Y, Z) = 4$ біти.

Але якщо порахувати щойно дану формулу, отримаємо

$$(3 + 2 + 1 - 1 - 1 - 1)\text{біти} = 3\,\text{біти} = \text{НІ!}$$[1]

Чому? Адже якщо між Z і Y існує взаємна інформація, а також між Z та Y, то частина цієї взаємної інформації може збігатися з тією, яку ми обчислимо між X та Y. У цьому випадку, наприклад, інформація про те, що X парне, дозволяє нам зробити висновок, що і Z парне, а інформація про те, що Z парне, — висновок, що й Y парне, але це та сама інформація, яку X дав би нам про Y. Ми двічі порахували частину наших знань, через що отримали занижену величину ентропію.

Правильна формула (як на мене) виглядає так:

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) - I(X;Z) - I(Z;Y) - I(X;Y|Z)$$

Тут останній член — $I(X;Y|Z)$ — означає: «інформація, яку X передає про Y за умови, що Z відоме». У такому разі X не дає нам ніякої інформації про Y за умови знання про Z, тому цей член дорівнює нулю — і рівняння дає правильну відповідь. Хіба не чудово?

«Ні, — скажете ви і матимете рацію, — бо ви не сказали мені, як порахувати $I(X;Y|Z)$ — лише навели словесне обґрунтування того, що це значення має дорівнювати нулю».

Обчислюємо $I(X;Y|Z)$ так, як ви очікували. $I(X;Y) = H(X) + H(Y) - H(X,Y)$, отже:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z).$$

1. https://youtu.be/tRVUOGUmxJI

Тепер, гадаю, вам кортить дізнатися, як рахувати умовну ентропію. Початкова формула ентропії має такий вигляд:

$$H(S) = Sum i\colon p(S_i) \times (-log_2(p(S_i))).$$

Якщо після цього ми дізнаємося про новий результат Z_0, рівень нашої невизначеності щодо S буде таким:

$$H(S|Z_0) = Sum i\colon p(S_i|Z_0) \times (-log_2(p(S_i|Z_0))).$$

Якщо ми хочемо дізнатися новий результат Z, але ще не знаємо, яке саме це Z, то в середньому очікуємо, що опісля рівень нашої невизначеності щодо S становитиме приблизно:

$$H(S|Z) = Sum j\colon (p(Z_j) \times Sum i\colon p(S_i|Z_j) \times (-log_2(p(S_i|Z_j))))$$

Ось так рахуються умовні ентропії, з яких, своєю чергою, ми можемо отримати умовну взаємну інформацію.

Звідси випливають *різноманітні* допоміжні теореми, наприклад:

$$H(X|Y) = H(X,Y) - H(Y)$$

та

якщо $I(X;Z) = 0$ і $I(Y;X|Z) = 0$, тоді $I(X;Y) = 0$,

але зараз не про це.

«Але ж, — спитаєте ви, — як це стосується природи слів і їхніх прихованих баєсових структур?».

Який же я *невимовно* радий, що ви поставили це питання, бо ж я якраз думав розказати вам, але не знав, сподобається вам чи ні. Але спершу ще декілька попередніх зауваг.

Ви запам'ятаєте — так, *запам'ятаєте*, — що між взаємною інформацією та баєсовим свідченням існує двоїстість. Взаємна інформація позитивна, якщо (і тільки якщо) ймовірність бодай якихось спільних подій $P(x, y)$ не дорівнює добутку

ймовірностей окремих подій $P(x) \times P(y)$. Зі свого боку, це положення відповідає умові, що між X та Y баєсове свідчення існує:

$$I(X;Y) > 0 = > \quad P(x,y)! = P(x) \times P(y) \quad P(x,y)/P(y)! = P(x)$$
$$P(x|y)! = P(x)$$

Якщо ви використовуєте Z як умовну змінну, скоригуйте весь вивід відповідно:

$$I(X;Y|Z) > 0 = > \qquad P(x,y|z)! = P(x|z) \times P(y|z)$$
$$P(x,y|z)/P(y|z)! = P(x|z)$$
$$(P(x,y,z)/P(z))/(P(y,z)/P(z))! = P(x|z)$$
$$P(x,y,z)/P(y,z)! = P(x|z)\ P(x|y,z)! = P(x|z)$$

Останній рядок виводу звучить так: «Навіть якщо ми знаємо Z, інформація про Y все ще змінює переконання про X».

І навпаки — як у нашому початковому випадку, де Z парне чи непарне, Z відокремлює X від Y: якщо знаємо, що Z парне, інформація про те, що Y у четвертому стані, *нічого не говорить* нам про те, чи X у другому, четвертому, шостому чи восьмому стані. Або якщо ми знаємо, що Z непарне, то інформація про те, що X у п'ятому стані, не дає нам нічого про те, чи Y у першому чи третьому. З набуттям інформації про Z, X та Y стали *умовно незалежними*.

Умовна незалежність — надзвичайно важливе поняття у теорії ймовірностей: наприклад, без умовної незалежності всесвіт не мав би структури.

Однак, тут я налаштований розповісти лише про один конкретний тип умовної незалежності — про випадок, коли центральна змінна унезалежнює одні змінні від інших, що оточують її, немов центральне тіло з мацаками.

Нехай у нас є п'ять змінних: U, V, W, X, Y. Ба більше, скажімо, для кожної пари цих змінних одна змінна — одне свідчення щодо іншої. Оберете, наприклад, U та W — інформація про те, що $U = U_1$, дасть вам щось таке, чого до цього ви не знали про ймовірність $W = W_1$.

Непосильний хаос у висновках? Свідчення вийшли з-під контролю? Не обов'язково.

Можливо, *U* означає «говорить людською мовою», *V* — «має дві руки та по десять пальців на кожній», *W* — «носить одяг», *X* — «вразливий до болиголова», а *Y* — «червона кров». Отже, якщо у дійсному світі вам трапиться якийсь об'єкт — хай то буде яблуко чи камінь — і ви дізнаєтеся, що він розмовляє китайською, ви, певно, надасте набагато вищу ймовірність того, що він носить одяг; а якщо дізнаєтеся, що цей об'єкт не вразливий до болиголова, надасте дещо меншу ймовірність, що у нього червона кров.

Отже, якісь із цих правил сильніші за інші. Маємо випадок Фреда, в якого бракує одного пальця через інцидент з вулканом, немовля Берні, який ще не заговорив, та IRC-бота Ірвінґа, що видає речення, але в нього не тече кров. Тож, якщо ми дізнаємося, що певний об'єкт не носить одягу, це свідчення не виключає того, що його мовленнєві здібності можуть розказати нам про колір його крові. Якщо цей об'єкт не носить одягу, але *говорить*, можливо, перед нами нудистка Неллі.

Від того цей випадок іще цікавіший за п'ять цілочисельних змінних, котрі всі парні або всі непарні, але в іншому нескорельовані. У такому випадку знання про *будь-яку* з тих змінних виключає все, що інформація про другу змінну може дати нам про третю.

А ось тут у нас є залежності, що не зникають, щойно ми дізнаємося про одну змінну, як показує випадок з нудисткою Неллі. Чи ця обтяжливість із виведенням висновків непереборна?

Не бійтеся! Адже може існувати ще якась шоста змінна *Z*, і якби ми її знали, вона справді унезалежнила *б* кожну пару змінних одна від одної. Може існувати деяка змінна *Z* — навіть якщо нам доведеться *сконструювати Z*, а не спостерегти її напряму — така, що:

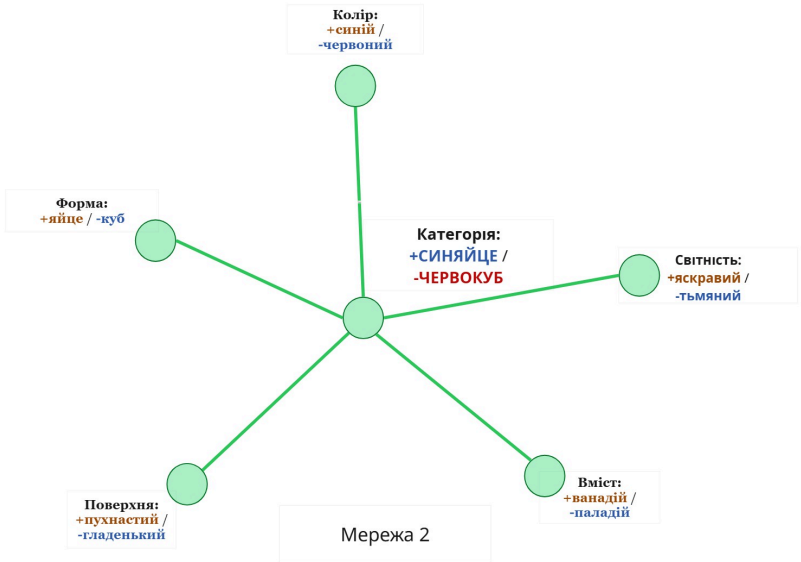
$p(u|v, w, x, y, z) = p(u|z)$ $\quad$ $p(v|u, w, x, y, z) = p(v|z)$
$p(w|u, v, x, y, z) = p(w|z)$ ...

Певно, *за умови*, що об'єкт — «людина», імовірності того, що вона говорить, носить одяг і має стандартну кількість пальців, незалежні. У Фреда може не бути пальця, але від того він не стає більшим нудистом, ніж людина поряд. Нудистка Неллі ніколи не носить одягу, але ця інформація не впливає на ймовірність того, що вона розмовляє. Немовля Берні ще не розмовляє, але всі його кінцівки на місці.

Цей метод називається «наївним баєсовим класифікатором», бо припущення в його основі не зовсім істинне, але ми *вдаємо*, що воно таке, бо так можемо спростити собі життя і розрахунки. Ми не відстежуємо окремо вплив носіння одягу на здатність до мовлення за умови числа пальців. Ми просто використовуємо спостережувану інформацію, щоб слідкувати за ймовірністю того, що той об'єкт — людина (або щось типу шимпанзе чи робота), а потім за допомогою наших переконань про центральний клас передбачаємо все, чого ще не побачили, — вразливість до болиголова, наприклад.

Будь-яке спостереження за змінними U, V, W, X та Y слугують лише свідченням для змінної Z з центрального класу. Опісля ми використовуємо апостеріорний розподіл по Z для формування будь-яких необхідних прогнозів щодо неспостережуваних змінних U, V, W, X та Y.

Звучить знайомо? А має:



По суті, якщо використовувати відповідні типи блоків нейронної мережі, ця «нейромережа» виявиться *математично* еквівалентною методу наївного баєсового класифікатора. Центральному блоку достатньо лише логістичного порогу — S-подібної кривої реакції, — а ваги вхідних даних мають відповідати логарифмам відношень правдоподібностей тощо. Насправді цілком імовірно, що саме у цьому причина частої ефективності логістичної реакції в нейронних мережах: непомітно для розробників вона дозволяє застосовувати елементи баєсового міркування.

Тільки тому, що хтось пропонує вам алгоритм, називаючи його «нейронною мережею» і обсипаючи його модними словечками на кшталт «неохайний» та «емерджентний», і при цьому гордо заявляє, що гадки не має, як працює ця натренована мережа, не розраховуйте на те, що їхній невеличкий алгоритм ШІ *дійсно* виходить за межі логіки. Адже ця модель тимчасових рішень, якщо вона робоча, виявиться баєсівськи

структурованою — вона може навіть виявитися еквівалентною алгоритму того типу, що зветься «баєсовим».

Навіть якщо на поверхні вона не *виглядає* як баєсова.

Тоді ви *знаєте*, що баєсіанці просто почнуть пояснювати, як саме працює цей алгоритм, які базові припущення він відображає, якими закономірностями середовища керується, де працює, а де лягає, і навіть надаватимуть зрозуміле значення ва́гам, яких навчилася мережа.

Розчаровані? Чи ні?

Слова — руків'я пензлів думки

(Нам уже час закруглятися з математично навантаженими есеями. Принаймні поки. Даруйте, якщо, як не дивно, на день-два я буду змушений вдатися до РаціоЦитат. Я зараз беру участь у конференції AGI-08, і, як на першу сесію, вона не настільки погана, як я боявся.)

От я скажу вам: «Ну й дивина: вперше в готелі бачу лампи з трикутними світильниками».

Можливо, ви вже уявили собі цю картину, а можливо, ще ні (якщо ні, уявіть зараз). Яким вам в голові бачиться «трикутний світильник»?

Скло було гострої чи заокругленої форми?

Коли словосполучення «трикутний світильник» вперше спало мені на думку (ні, в готелях таких немає), внутрішній взір показав мені пірамідальний світильник з гострими кутами, а потім мій розум створив альтернативну версію — петлю флюоресцентної лампочки форми гладенького трикутника.

Наскільки я можу судити, ніяких свідомих чи вербальних думок я не задіював — лише безсловесна рефлекторна реакція на уявне зображення гострого скла, проблема дизайну якого розв'язувалася ще до того, як я навіть встиг сформулювати її словами.

Вірте чи ні, а кілька десятиліть точилися неабиякі суперечки стосовно того, чи справді люди могли уявляти образи у себе в голові — власне *картинка* стільця, — чи вони просто наївно *гадали*, що могли (а з пантелику їх збивала власна інтроспекція — дуже погана і заборонена діяльність), тоді як в їхньому мозку активувався маленький ярлик «стілець» — такий собі ліспток.

З усіх сил тримаюся, щоб не сказати: «Яка ж феноменальна дурість», — позаяк завжди слід пам'ятати про ефект ретроспективи. Але: яка ж все-таки феноменальна дурість!

Гадаю, ця академічна парадигма здебільшого була нездоровою спадщиною біхевіоризму, в якому заперечували наявність думок у людей і намагалися пояснити усі людські прояви «рефлексами», зокрема й мовлення. Можливо, біхевіоризм певною мірою заслуговує на власний допис, оскільки він був таким собі збоченням раціоналізму. Але зараз не про це.

«А чому ви називаєте це "дурістю"? — запитаєте ви. — Звідки вам *знати*, що ваш мозок відтворює візуальні образи? Просто тому, що ви закриваєте очі і бачите їх?».

Раніше, коли довкола цього питання точилися суперечки, на нього було складніше відповісти. Якщо ви хотіли «науково», а не інтроспекційно, довести існування мисливих образів, вам довелося б зробити висновок про їхнє існування на основі таких експериментів, як, наприклад, показати піддослідним два об'єкти і запитати, чи можна один повернути так, щоб він збігся з іншим. Час реакції лінійно пропорційний куту, на який треба повернути. Таке спостереження легко пояснити, якщо ви саме уявляєте образ і безперервно обертаєте його з постійною швидкістю, важко ж — якщо ви перевіряєте символічні риси того образу.

Сьогодні ж ми можемо нейровізуалізувати ці маленькі картинки у зоровій корі. Тому так — ваш мозок дійсно відображає деталізоване зображення того, що він бачить чи уявляє. З цього приводу раджу ознайомитися з працею Стівена Кóссліна «Зображення й мозок: результат дебатів про óбрази» (Image and Brain: The Resolution of the Imagery Debate).

Люди мають складнощі зі словами почасти через те, що вони навіть не усвідомлюють, яка складність ховається за ними.

Чи можете ви уявити «зеленого собаку»? А «сирне яблуко»?

«Яблуко» — не просто послідовність трьох складів чи шести літер. Власне, «яблуко» — це тінь, кінчик хвоста тигра.

Слова, чи радше поняття, — це пензлі: ними ви можете вималювати óбрази зі свого розуму. Буквально вималювати, якщо поняттями ви створюєте зображення у своїй зоровій корі. А за допомогою спільних позначень ви можете дістатися до чийогось

розуму, взяти їхні пензлі і намалювати о́брази в *їхньому* розумі, наприклад — накидати у зоровій корі обриси зеленого собаки.

Але не подумайте: склади, що ви їх посилаєте через повітря, літери чи інтернет, — не ті самі склади, якими малюють зображення у зоровій корі. Для зображень потрібні деякі складні інструкції, що не вписуються в послідовність літер. «Яблуко» займає п'ять байтів, а намалювати яблуко з нуля коштуватиме і того більше даних.

«Яблуко» — просто теґ, прикріплений до істинного й безсловесного поняття *яблука*, котрий може намалювати зображення у вашій зоровій корі або зіткнутися з «сиром», або розпізнати яблуко, коли ви його бачите, або дозволити відтворити смак його архетипу у пирозі з яблуками, або навіть запустити моторні рухи для поїдання яблука...

І йдеться не про згадування зображення, а про дещо складніше. А інакше — як вам вдалося вимислити таку комбінацію, як «трикутний світильник» — накласти трикутність на світильники, не змінюючи суті обидвох, навіть якщо ніколи в житті не бачили нічого подібного?

Не повторюйте помилок біхевіористів. Мова не зводиться лише до звуку в повітрі. Ці ярлики — лише покажчики: «дивіться в зоні пам'яті 1387540». Рано чи пізно, коли вам передають покажчик, потрібно розібрати його і подивитися у ту зону 1387540.

На що ж вказує слово?

Хиби прихованої змінної

> Альберт: Щоразу, коли я чув, як падає дерево, воно видавало звук. Відповідно, решта дерев при падінні так само видає звуки. Не думаю, що щось зміниться, якщо мене там не буде. Беррі: Зачекай. А як воно може видавати звук, якщо ніхто не слухає?

Прописуючи діалог між Альбертом та Беррі про те, чи видає в безлюдному лісі падаюче дерево звук, я іноді ловив себе на думці, що перестаю розуміти своїх героїв. Я почав втрачати внутрішнє відчуття, чому *хтось* узагалі наводить такі аргументи, хоча неодноразово був свідком таких суперечок.

У таких випадках я повторював про себе: «Або при падінні дерево видає звук, або ні» — щоб відновити запозичене відчуття обурення.

Евристика (P або $\neg P$) не завжди надійна, якщо замість P ви підставите довільне речення українською. Речення «Це речення хибне» не можна послідовно розглядати як істинне або хибне. Тут постає старе класичне провокаційне питання — «Ви все ще б'єте свою дружину?».

Якщо ви математик і вірите у класичну (а не інтуїціоністську) логіку, можете продовжувати наполягати на тому, що (P чи $\neg P$) — теорема: наприклад, сказати, що «Це речення хибне» взагалі не є реченням.

Але такі розв'язки витончені, і їх достатньо, щоб проілюструвати необхідність витонченого підходу. Не можна щоразу просто наполягати: «Або так, або ні!».

Отже, чи при падінні дерево видає звук чи не видає, чи... ?

Певна річ, $2 + 2 = X$. Чи не дорівнює? Ну, можливо, *якщо* це *справді* те саме X, те саме 2 і ті самі $+$ та $=$. Якщо в деяких випадках X дорівнює п'яти, а в інших — чотирьом, ваше обурення може бути невиправданим.

Взагалі, щоб можна було стверджувати, що (P чи $\neg P$) має бути обов'язковою істиною, символ P має означати *точно* те саме в обох частинах дилеми. «Або падіння спричиняє звук, або ні!» — але якщо Альберт: : звук не той самий, що Беррі: : звук, нічого парадоксального у тому, що дерево видає Альберт: : звук, а не Беррі: : звук.

(Ідіому : : я взяв з мови C++, щоб уникнути конфліктів іменних просторів. Якщо у вас два різні пакети, що визначають клас Звук, ви можете написати Пакет1: : Звук, щоб вказати, котрий саме Звук ви маєте на увазі. Ця ідіома, гадаю, не широко відома. А шкода, бо на письмі мені її часто бракує.)

Ця відмінність може бути ледь помітною: Альберт і Беррі можуть упевнитися, що дерево те саме у тому самому лісі, той самий випадок падіння, і все це для того, щоби пересвідчитися, що між ними дійсно існує суттєва розбіжність з приводу саме цієї події. А потім вони забувають звірити, чи співвідносять цю подію саме із тим самим поняттям.

От узяти продуктовий магазин, в який ви ходите найчастіше: він розташований на лівій стороні вулиці чи правій? Звісно, немає ніякої *визначеної* «лівої сторони» вулиці — тільки *ваша* ліва сторона, коли ви йдете по ній з певного напрямку. Велика частина нашого слововжитку виступає функціями прихованих змінних, обумовлених контекстом.

Розв'язувати таку задачу — явище, знане як «дейксис мовця» — у програмі штучного інтелекту, призначеній для аналізу мови, це справжня му́ка, що потребує достобіса зусиль.

«Мартін сказав Бобу, що будівля ліворуч від нього». Але «ліворуч» — слово-функція, значення якого визначається змінною, що залежить від мовця і непомітно витягується з сусіднього контексту. Чиє «ліворуч» мається на увазі — Бобове чи Мартінове?

Змінні у хибі зі змінною питання часто позначені не так чітко — не так просто, як-от: «Слухай, а ти не думаєш, що $Z + 2$ дорівнює 6?»

Якщо конфлікт іменних просторів вводить два різні поняття, що виглядають як «одне й те саме поняття» через однакову

назву, — або через стиснення мапи вводить дві різні події, що виглядають як одна, бо для них не завели окремих ментальних файлів; або та сама функція обчислюється по-різному в різних контекстах, — тоді сама реальність стає мінливою і плинною. Принаймні саме таким алгоритм відчувається зсередини. Адже ваш внутрішній зір бачить мапу, а не територію безпосередньо.

Якщо у вас є питання із прихованою змінною, значення якої в різних контекстах обчислюється по-різному, *складається враження*, ніби сама дійсність не стабільна: те, що бачить ваш внутрішній зір, змінюється залежно від того, куди він спрямований.

Таке часто спантеличує студентів (і професорів-постмодерністів), які відкривають для себе багатозначні речення, а думають, що відкрили нестабільну частину дійсності.

«Боже правий! "Сонце обертається довкола Землі" істинне для Гунґи з племені мисливців-збирачів, а для космонавтки Амари хибне! Не існує ніякої незмінної істини!». Вправу деконструювати цю бздуру другокурсника залишу читачеві.

А втім, навіть я спочатку ловив себе на тому, як писав: «Якщо у деяких випадках X дорівнює п'яти, а в інших — чотирьом, твердження "$2 + 2 = X$" може не мати визначеного значення істинності». Не існує *якогось* речення зі *змінним* значенням істинності. У твердженні «$2 + 2 = X$» *немає* значення істинності. Воно не є *твердженням* — у тому сенсі, в якому математики визначають саме поняття «твердженнєвості» — так само, як не є твердженням «$2 + 2 =$ », а «Фред перестрибнув через» не є граматичним реченням.

Та ця хиба може просочуватися в розум навіть тоді, коли ви нібито знаєте краще, бо саме так алгоритм відчувається зсередини.

Тридцять сім способів схибити зі словом

Дехто з читачів обов'язково скаже, що «Тридцять сім способів нерозумного слововжитку» або «Тридцять сім способів неоптимального використання категорій можуть мати негативні побічні ефекти на ваше мислення» — краща назва для цього есею.

З цього довжелезного переліку варто винести один з основних уроків: положення «ні в якому разі мій вибір X не може бути "хибним"» на практиці майже завжди помилкове, хай там що у тій теорії. Ви завжди можете помилятися. Навіть якщо помилятися теоретично неможливо, ви все ще можете помилятися. Хай би чим ви займалися, у вас немає способу уникнути провалів. Таке життя.

До того ж я можу надати слову «хибний» будь-яке значення, яке мені тільки заманеться, — слово ж не може бути *хибним*.

Особисто я вважаю, що слово «хибний» цілком виправдано вживати у таких випадках:

1. *Насамперед слово не в'яжеться з дійсністю.* Сократ — фремстер? Так чи ні? («Притча про кинджал»)
2. *Ваш аргумент, якщо він спрацював, зміг примусити дійсність стати інакшою, і все завдяки вибору іншого визначення.* Сократ — людина, а люди, за визначенням, смертні. То якщо вважати людей безсмертними, Сократ житиме вічно? («Притча про болиголов»)
3. *Ви намагаєтеся встановити істинність будь-якого емпіричного твердження «за визначенням».* Сократ — людина, а люди, за визначенням, смертні. То чи буде *логічною істиною* наше емпіричне передбачення, що Сократ звалиться з ніг, якщо вип'є настоянку болиголова? Складається враження, ніби існують логічно можливі

позбавлені внутрішніх суперечностей світи, в яких Сократ не падає з ніг, бо, скажімо, завдяки якійсь біохімічній мутації має імунітет до болиголова. Логічні істини істинні в усіх можливих світах, а отже, ніколи не вказують на те, *в котрому* з можливих світів живете ви. А все, що можна встановити «за визначенням», є логічною істиною. («Притча про болиголов»)

4. *Ви несвідомо навішуєте загальноприйняту етикетку на якийсь об'єкт і при цьому не користуєтеся тим самим словесним визначенням.* Ви чудово знаєте, що Боб — «людина», хоча за вашим визначенням так ніколи й не зможете назвати Боба «людиною», доки не побачите, як він помре. («Притча про болиголов»)
5. *Даючи чомусь слово, ви маскуєте індуктивний висновок, який лежить в основі цього означення і який може виявитися хибним або бути оскарженим.* Якщо останні одинадцять яйцеподібних об'єктів, що ви їх дістали з контейнера, були синіми, а останні вісім кубів — червоними, за індукцією ви маєте сказати, що це правило діятиме і в майбутньому. Але якщо сині яйця ви називаєте «синяйцями», а червоні куби — «червокубами», наступного разу, коли опустите руку в контейнер і намацаєте яйце, закономірно подумаєте: «О, синяйце!». («Слова — приховані умовиводи»)
6. *Ви намагаєтеся дати визначення слову словами і зрештою доходите до ще абстрактніших слів і не можете підібрати до них прикладів.* «Що таке червоний? — Червоний — це колір. — Що таке колір? — Колір — це властивість об'єкта. — Що таке об'єкт? А властивість?». І вам не спаде на думку вказати на яблуко чи знак «Проїзд без зупинки заборонено». («Обсяг та зміст поняття»)
7. *Обсяг поняття не збігається з його змістом.* Ми не усвідомлюємо, як ідентифікуємо червону цятку в небі як «Марс», і цей автоматизм відбувається, певно, незалежно від

того, чи спробуєте ви визначити «Марс» як «бога війни». («Обсяг та зміст поняття»)

8. *Ваше словесне визначення не покриває і крихітної частини спільних властивостей категорії, але ви намагаєтеся думати, що покриває.* Коли філософи з Академії Платона проголосили, що найкращим визначенням людини є «двоноге без пір'я», кажуть, що Діоген Кінік приніс обскубану курку і привселюдно назвав її «платонівською людиною». Платоністи ж відразу змінили своє визначення на «двонога істота без пір'я з широкими нігтями». («Скупчення подібностей»)
9. *Ви намагаєтеся оцінювати приналежність до категорії за принципом «належить/не належить» і нехтуєте більш-менш типовими підгрупами.* На відміну від вільшанок з голубами, качки та пінгвіни менш типові птахи. Як показують міжгрупові експерименти, на думку піддослідних, на острові хвороба поширюватиметься з більшою ймовірністю від вільшанок до качок, аніж навпаки. («Типовість та асиметрична подібність»)
10. *На практиці словесне визначення досить ефективно вказує на цільове скупчення подібних об'єктів, однак ви прискіпуєтеся до виняткiв.* Не кожна людина має десять пальців чи носить одяг, чи користується мовою. Але якщо ви шукаєте емпіричні скупчення об'єктів, об'єднані цими властивостями, випадкова людина з дев'ятьма пальцями не зіб'є вас із пантелику. («Структура скупчень у просторі можливих об'єктів»)
11. *Ви питаєте, чи щось «є» чи «не є» членом категорії, але не можете сформулювати питання, на яке хочете почути відповідь.* Що таке «людина»? Немовля Берні — «людина»? «Правильна» відповідь може значною мірою залежати від конкретного запитання, на яке ви *справді* хочете відповіді: «Чи можна Берні вигодовувати болиголовом?» або «Чи буде Берні хорошим чоловіком?». («Замасковані запити»)

12. *До інтуїтивно сприйнятих ієрархічних категорій ви ставитеся як до єдино правильного способу розбирати світ на частини, не усвідомлюючи, що можливі й інші форми статистичного висновування — навіть попри те, що ваш мозок ними не користується. Людині* легше побачити «синяйцевість» або «червокубність» об'єкта, ніж помітити, що червоні об'єкти не світяться в темряві, а червоні пухнасті об'єкти мають усі інші ознаки синяйця. Інші статистичні алгоритми працюють по-іншому. («Нейронні категорії»)
13. *Ви говорите про категорії так, ніби вони дані людям з платонівського світу, а не сформовані на основі висновків, виведених фізичним мозком.* Античні філософи казали: «Сократ — людина», — а не: «На основі чуттів мій мозок відносить Сократа до поняття "людина"». («Як алгоритм відчувається зсередини»)
14. *Ви сперечаєтеся про приналежність до категорії навіть після того, як відсіяли усі питання, що могли б залежати від висновування на основі категорій.* Ви побачили, що об'єкт синій, яйцеподібний, пухнастий, гнучкий, матовий, світиться у темряві і має всередині паладій. Які ще питання *лишилися* у суперечці «Це синяйце»? Якщо ваша нейромережа, відповідальна за категоризацію, має (метафоричний) центральний блок, завдяки якому ви висновуєте синяйцевість, вас може не покидати *відчуття*, що лишилися ще якісь питання без відповіді. («Як алгоритм відчувається зсередини»)
15. *Ви дозволяєте суперечці зіслизнути до обговорення визначень, хоча на початку не мали такого наміру.* Якби перед суперечкою про дерево у безлюдному лісі ви запитали у двох майбутніх сперечальників, чи, на їхню думку, «звук» слід визначати як «акустичні коливання» чи як «слуховий досвід», імовірно, вони б сказали підкинути монетку. Питання визначення слова набуває політичного забарвлення тільки у розпалі суперечки. («Суперечка про визначення»)

16. *Ви вважаєте, що слова мають значення, причому ті значення — властивість самих слів, а не створені мозком асоціації між якимось позначенням і певним поняттям.* Коли хтось кричить: «Яаааааай! ТИГР!», — еволюція не помилує організм, який почне роздумувати: «Хмм... я почув склад [тигр], котрий мої одноплемінники асоціюють зі своїми внутрішніми аналогами мого уявлення про *тигра* і які *аяяяяаааааай* ХРУСТЬ ХРУСТЬ КОВТЬ». Тому мозок обирає коротший шлях і, схоже, вважає значення тигровості властивістю самого позначення. А сучасні люди вже сперечаються про *правильне значення* слова «звук». («Відчути значення»)
17. *Ви не перестаєте сперечатися про значення слова навіть після того, як усі сторони чудово зрозуміли одна одну.* Здатність людини прикріпляти позначення до понять — це засіб спілкування. Коли люди *хочуть* спілкуватися, нас важко зупинити: якщо немає спільної мови, ми почнемо малювати символи на піску. Коли кожен з вас розуміє, що на думці в іншого, вітаю: комунікація *відбулася*. («Аргумент від поширеного вжитку»)
18. *Посеред емпіричної чи моральної суперечки ви дістаєте словник.* Укладачі словників фіксують мововжиток, а не регулюють його. Якщо поширене визначення проблемне — якщо «Марс» тлумачать як бога війни, «дельфіна» — як короля риб, а «негра» — як окрему від людей категорію, — словник повторюватиме цю стандартну помилку. («Аргумент від поширеного вжитку»)
19. *Ви дістаєте словник посеред кожної суперечки.* Серйозно: що в біса змушує вас думати, що лексикографи мають право визначати приналежність «атеїзму» до «релігії» чи що-небудь іще? Якщо на кону якесь важливе питання, ви справді думаєте, що лексикографи мають доступ до тієї остаточної мудрості, котра допоможе покласти край суперечці? («Аргумент від поширеного вжитку»)

20. *Ви без причини ігноруєте поширений слововжиток і лише ускладнюєте іншим завдання зрозуміти вас.* Швидкорозчинний плутоній з бубликом без ручок. («Аргумент від поширеного вжитку»)
21. *Ви використовуєте складні перейменування, щоб створити ілюзію висновування.* Визначаєте «людину» як «смертне двоноге без пір'я»? То так і напишіть: «Усі [смертні двоногі без пір'я] смертні; Сократ — [смертне двоноге без пір'я]; отже, Сократ смертний». Так уже не вражає. Еге ж? («Порожні позначення»)
22. *Ви вплутуєтеся в суперечки, яких могли б уникнути, якби не вживали того слова.* Якщо Альбертові і Беррі не дозволено вимовляти слово «звук», тоді Альберт буде змушений сказати: «При падінні у безлюдному лісі дерево видає акустичні коливання», — а Беррі: «При падінні у безлюдному лісі дерево не породжує ніякого слухового досвіду». Коли саме слово створює проблему, тоді найпростіше рішення — прибрати його і пов'язані з ним синоніми. («Затабуюйте це слово»)
23. *Через компактне слівце ви не бачите подробиць того, про що намагаєтеся думати.* Що насправді відбувається у школах, якщо не називати це «освітою»? Що таке диплом, якщо не називати це «дипломом»? Якщо монетка випала числом, яке її радіальне положення? Що таке «істина», якщо не казати «правильний», «точний», «відображати», «віддзеркалювати», «семантичний», «вірити», «знання», «мапа», «дійсний» чи будь-який інший простий термін? («Замініть символ суттю»)
24. *У вас є тільки одне слово, але два чи більше різних об'єктів у дійсності, тому вся інформація про них скидається в один нерозрізнюваний ментальний кошик.* У звичні обов'язки слідчого входить спостерігати, якого кольору сукню носила Керол минулої ночі чи якого кольору у неї волосся, а також припускати, що Керол фарбує волосся. Але тільки

кмітливіший слідчий припустить, що існують дві Керол: можливо, носили червону сукню і ходили з чорним волоссям дві різні людини. («Помилки стиснення»)

25. *Ви бачите закономірності там, де їх немає, і збираєте інші властивості з ваших висновків навіть тоді, коли вздовж того виміру немає подібностей.* У Японії вірять, що люди з групою крові А щирі й креативні, з групою В — гарячкуваті й добродушні, з групою О — згідливі та соціальні, а з групою АВ — холодні та стримані. («Категоризація має наслідки»)
26. *Ви намагаєтеся додати до значення слова конотації, яких немає в словнику, і висувати подальші твердження з цього визначення.* У словнику «вігґін» тлумачиться як людина із зеленими очима та чорним волоссям. Але також слово «вігґін» передає конотацію, пов'язану зі злочинністю і шкодою білченятам, але цього у словнику вже не записано. Ви показуєте на когось пальцем: «Зелені очі, чорне волосся. Бачиш: я ж казав, що він вігґін! Ось зараз іще срібні ложки покраде». («Приховані конотації»)
27. *Ви стверджуєте: «X, за визначенням, — Y!». У таких випадках ви майже завжди намагаєтеся додати конотацію Y, якої не було у наданому визначенні.* Ви тлумачите «людину» як «двоноге без пір'я», показуєте на Сократа: «Ані пір'їнки, дві ноги — точно людина!». Але ж *насправді* вас цікавить смертність. Якби предметом спору було число ніг у Сократа, сперечальник відповів би: «В сенсі у Сократа дві ноги? Ми ж про це сперечаємося!». («Аргумент "За визначенням"»)
28. *Ви стверджуєте: «P, за визначенням, — Q!».* Якщо десь у полі ви побачили Сократа з біологами, які ходили збирати зілля-протидію до болиголова, з того, що ви будете заходитися: «Люди, за визначенням, смертні!» — діла не буде. Найчастіше ви відчуваєте потребу затягнути гайки, наполягаючи на істинності чогось «за визначенням», коли

з'являється інформація, що ставить автоматичний висновок під сумнів. («Аргумент "За визначенням"»)

29. *Ви намагаєтеся встановити приналежність до емпіричної сукупності «за визначенням».* У вас немає потреби казати: «*За визначенням*, індуїзм — релігія!» — бо кожному і так зрозуміло, що індуїзм — релігія. І не просто *«за визначенням»*, а *насправді*. Атеїзм не нагадує центральних членів скупчення «релігія», тож якби не те, що атеїзм — це релігія *за визначенням*, ви, певно, подумали б, що атеїзм *не є* релігією. Ось тому вам доводиться розбивати опозицію ствердженням, що «Атеїзм — релігія» — істинне *за визначенням*, бо ніяким іншим чином істинним воно не є. («Аргумент "За визначенням"»)
30. *Своїм визначенням ви проводите межу довкола об'єктів, що насправді не схожі між собою.* Можете заявити (якщо хочете), що до «риб» *належать* лосось, гуппі, акула, дельфін та форель, але не медузи й водорості. Можете заявити (якщо хочете), що це просто перелік, а перелік не може бути «хибним». Або можете перестати корчити з себе дурника і визнати, що дельфіни не належать до риб. («Де ж провести межу?»)
31. *Коротким словом ви позначаєте щось таке, що не згадуєте надто часто, а довгим — таке, що згадуєте доволі часто. Такий вибір може вилитися в неефективне мислення чи навіть хибне застосування принципу Оккама, якщо ваш розум вважає, що короткі речення звучать «простіше».* Що звучить правдоподібніше: «Господь учинив диво» чи «Надприродна сутність, яка створила всесвіт, тимчасово призупинила дію законів фізики»? («Ентропія і короткі коди»)
32. *Ви проводите межу довкола об'єму простору, щільність якого не більша за звичайну, себто у вас виходить так, що асоційоване слово не відповідає жодним виконуваним баєсовим висновкам.* Навіщо нам слово «вігґін», якщо

зеленоокі люди не частіше за інших мають чорне волосся (чи навпаки) і вони не поділяють жодних інших ознак? («Взаємна інформація та щільність у просторі об'єктів»)

33. *Ви проводите непросту межу без якоїсь причини її проводити.* Визначати слово так, щоб посилатися на всіх людей, окрім чорношкірих, дещо підозріло. Якщо ви не даєте підстав проводити конкретно цю межу і намагаєтеся утворити «довільне» слово для цієї території, то поводитеся як той слідчий: «Що ж, так чи інакше у мене немає аніяковісінької зачіпки, хто міг вбити тих сиріт... а чи брали ми до уваги Джона К. Віффлгайма з Норкл-роуд 1234 як потенційного підозрюваного?». («Надекспоненційний простір понять та прості слова»)
34. *Ви використовуєте категоризацію для висновків про властивості, які не мають належної емпіричної структури — а саме умовної незалежності за умови знання класу, — необхідної для того, щоб наївний баєсів метод міг хоча б наближено оцінювати ці властивості.* Цей есей я не переповідатиму. Сходіть-почитайте самі. («Умовна незалежність та наївний баєсів метод»)
35. *Ви думаєте: слова — ніби ті маленькі ліспсимволи у вашому розумі, а не позначення, що функціонують як руків'я, якими ви скеровуєте складні пензлі думки і малюєте деталізовані зображення у вашій зоровій майстерні.* Уявіть «трикутний світильник». Що бачите? («Слова — руків'я пензлів думки»)
36. *Ви вживаєте слово з різними значеннями у різних контекстах так, ніби у кожному випадку воно означає те саме і, можливо, створює ілюзію чогось мінливого і плинного.* «Мартін сказав Бобу, що будівля ліворуч від нього». Але «ліворуч» — слово-функція, значення якого визначається змінною, що залежить від мовця і непомітно витягується з сусіднього контексту. Чиє «ліворуч» мається на увазі — Бобове чи Мартінове? («Хиби прихованої змінної»)

37. *Ви вважаєте, що визначення не можуть бути «хибними» або що «можна визначати слова як заманеться».* Таким ставленням ви вчитеся агресивно захищати власні минулі вчинки замість того, щоб звертати увагу на їхні наслідки чи визнавати свої помилки. («Тридцять сім способів неоптимального використання категорій можуть мати негативні побічні ефекти на ваше мислення»)

Усі ваші розумові процеси мають наслідки. Ваш мозок несвідомо працює на повну потужність і без вашого контролю.

Від переконання «Слова довільні, тому я можу визначати їх, як мені заманеться» стільки ж користі, скільки від їзди по тонкому льоду на повній швидкості. А ви ще й кажете: «Я дивлюся на кермо і не розумію, чим той радіальний кут такий особливий, — куди захочу, туди й кручу».

Якщо ви хочете куди-небудь дійти — або навіть просто вижити, — вам варто почати звертати увагу на ці кількадесят критеріїв оптимальності, що контролюють ваш слововжиток, визначення, категорії, класи, межі, позначення і поняття.

Вичерпний посібник з використання теореми Баєса

Правило, або теорема, Баєса — це закон імовірностей, що регулює *силу свідчень*, тобто *наскільки* переглядати нашу оцінку ймовірностей (наскільки змінювати думку), коли ми дізнаємося щось нове або отримуємо нове свідчення.

Вас може зацікавити теорема Баєса, якщо ви:

- фахівець, який у роботі послуговується статистикою (науковець чи лікар);
- програміст, який займається машинним навчанням;
- людина.

Вивчаючи «суперпрогнозистів» — людей, особливо вмілих у передбаченні подій майбутнього, — Філіп Тетлок помітив таке:

«Суперпрогнозисти — це люди, які вміють рахувати: багато хто з них знає про теорему Баєса і може застосувати її, якщо матиме потребу. Але вони рідко працюють із самими числами. Суперпрогнозистам важлива не стільки теорема, скільки її базовий принцип: поступово наближатися до істини шляхом постійного оновлення пропорційно вагомості свідчень». — «Суперпрогнозування», Філіп Тетлок і Ден Ґарднер[1]

1. Українською видана «Нашим Форматом»; переклала Ірина Мазарчук. — Прим. пер.

Геометричні візуалізації: знайомство з баєсовою статистикою

Баєсове міркування показує, як переглядати власні переконання залежно від свідчень.

А почнемо ми з розгляду одного сценарію, в якому сила свідчень виражається у кількісних показниках.

(Не хвилюйтеся, якщо не розумієте, як розв'язати наступну задачу. Невдовзі ми розберемо її.)

Скажімо, у вас, як у медсестри, є завдання провести обстеження групи учнів на захворювання під назвою дисизіт.[1]

- З минулих досліджень населення відомо, що у цю пору року на дисизіт хворіють близько 20% учнів.

Ви проводите тестування за допомогою лопатки для язика, що змінює колір: якщо вона чорніє, учень хворий на дисизіт.

- Серед пацієнтів із дисизітом у 90% лопатка забарвлюється у чорний.
- Та лопатка не завжди ідеально розпізнає хворобу і чорніє при контакті з 30% здорових учнів.

От учень заходить у медичний кабінет, проходить тестування, і його лопатка забарвлюється у чорний. Яка ймовірність того, що у нього дисизіт?

(*Якщо* думаєте, що розумієте, як розв'язати її, можете спробувати, а потім читати далі. Відповідь буде надана нижче.)

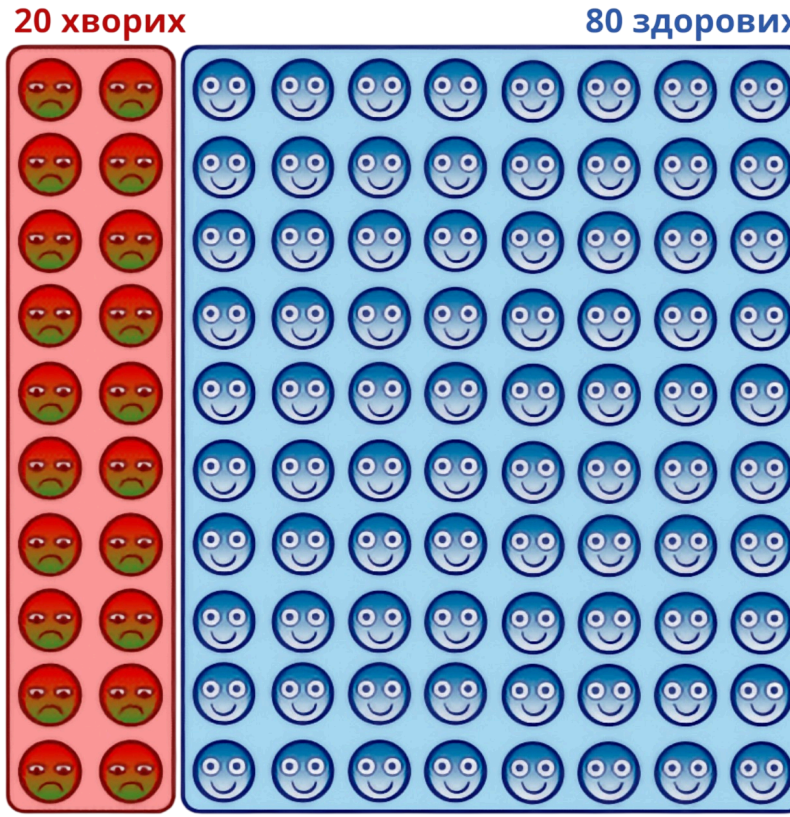
> Відповідь: Імовірність того, що лопатка чорніє при контакті з учнем, хворим на дисизіт, становить 3/7 — близько 43%.

1. Буквально «запалення, викликане хворобою».

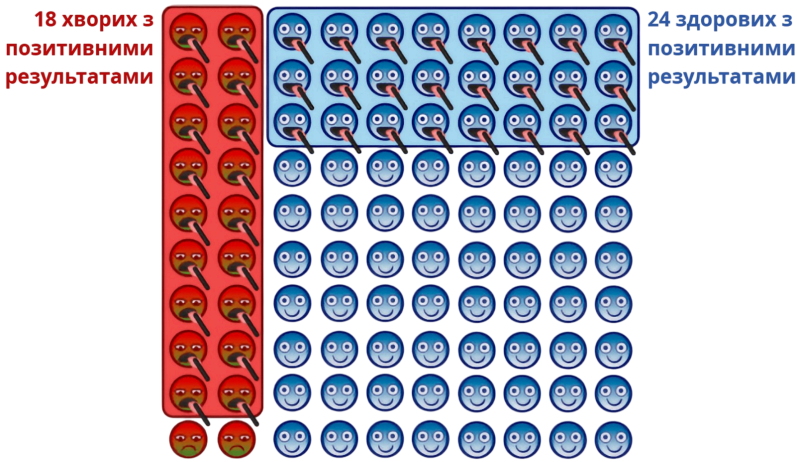
Цю задачу можна розв'язати складним способом або розумним і легким. Почнемо зі складного.

Спочатку уявляємо сукупність зі 100 учнів, з яких 20 хворі на дисизіт, а 80 здорові.[2]

2. Численні дослідження демонструють: якщо думати про точні числа способом «20 учнів зі 100» або «200 учнів з 1000», спонтанні роздуми про подібні задачі ймовірніше вестимуть у правильному напрямку, ніж роздуми над відсотками типу «20% учнів». Одне з них — Eddy DM. Probabilistic reasoning in clinical medicine: Problems and opportunities (https://pdf.stafforini.com/eddy-1982-probabilistic-reasoning-in.pdf). In: Kahneman D, Slovic P, Tversky A, eds. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press; 1982:249-267.



У 90% хворих учнів і в 30% здорових лопатка забарвлюється у чорний. Отже, ми спостерігаємо чорну лопатку у $90\% \times 20 = 18$ хворих учнів і в $30\% \times 80 = 24$ здорових.



Яка ймовірність того, що учень, у якого почорніла лопатка, хворий на дисизіт? На візуалізації видно, що з чорним індикатором 18 хворих учнів. Усього $18 + 24 = 42$ учні отримують позитивний результат при тестуванні. А тепер уявіть, що у вас є загальна група з учнями, у яких чорний індикатор, і ви вказуєте на випадкового учня. Який шанс того, що обраний учень хворий?



Остаточна відповідь: пацієнт із чорною лопаткою може виявитися хворим з імовірністю $18/42 = 3/7 = 43\%$.
Багатьом студентам-медикам, яким трапилася така задачка, відповідь виглядала контрінтуїтивною: тест правильно ж ідентифікує хворобу у 90% випадків! Якщо тест позитивний, чому ж тоді ймовірність того, що пацієнт хворий, менша за 50%? Водночас тест дає хибнопозитивний результат у 30% випадків. Як наслідок, у нас набагато більше хворих пацієнтів, аніж здорових.

Дійсно — цей тест дає нам якісь підстави вважати, що пацієнт хворий. Імовірність того, що він хворий, зростає з 20% до тесту до 43% після того, як ми побачили результат (чорна лопатка). Але з цього не можна виводити остаточний висновок — потрібно провести ще додаткові тести (можливо, дорожчі).

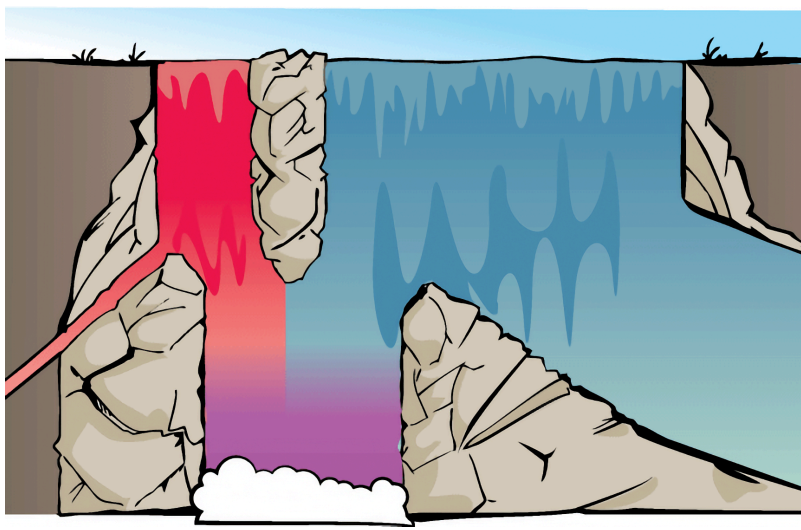
Якщо вам здається, що ви зрозуміли умови задачі, спробуйте відразу дати відповідь на наступне питання. Яка ймовірність того, що учень, в якого не чорніє лопатка (негативний результат), хворий на дисизіт? Знову ж — ми починаємо з 20% хворих і 80% здорових, при цьому негативний результат отримають 70% здорових учнів і тільки 10% хворих.

Відповідь: Уявіть 20 хворих учнів і 80 здорових. $10\% \times 20 = 2$ хворі учні отримають негативний результат. $70\% \times 80 = 56$ здорових учнів отримають негативний результат. Серед загальної кількості $2 + 56 = 58$ учнів з негативним результатом 2 учнів будуть хворими з негативним результатом. Отже, $2/58 = 1/29 = 3{,}4\%$ учнів з негативним результатом будуть хворими на дисизіт.

З розумним способом провести ті самі обчислення можете ознайомитися у наступному есеї «Каскадні візуалізації та співвідношення шансів».

Каскадні візуалізації та співвідношення шансів

Уявіть водоспад з двома верхніми руслами — червоним і синім. До уступу вони течуть окремо. На шляху до підніжжя якась частина води стікає в іншу сторону, а решта — у спільний нижній басейн.



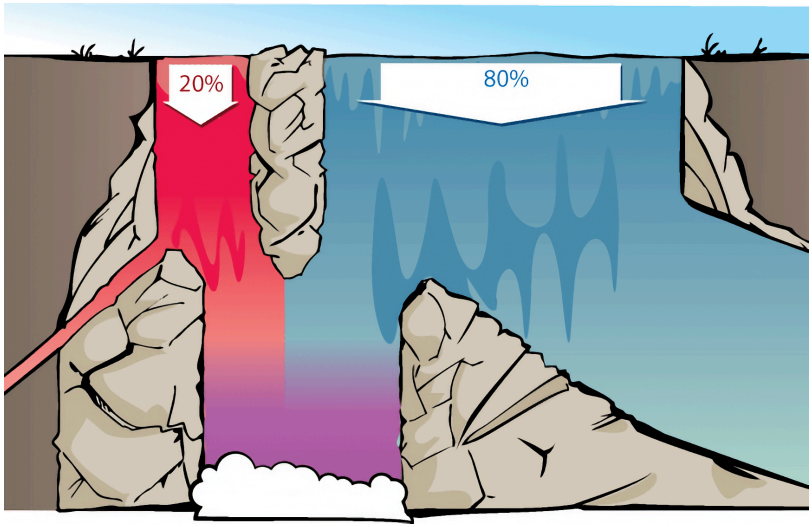
Припустімо, що

- з уступу водоспаду стікає 20 галонів червоної і 80 галонів[1] синьої води за секунду.
- 90% червоного і 30% синього русла стікають аж до дна.

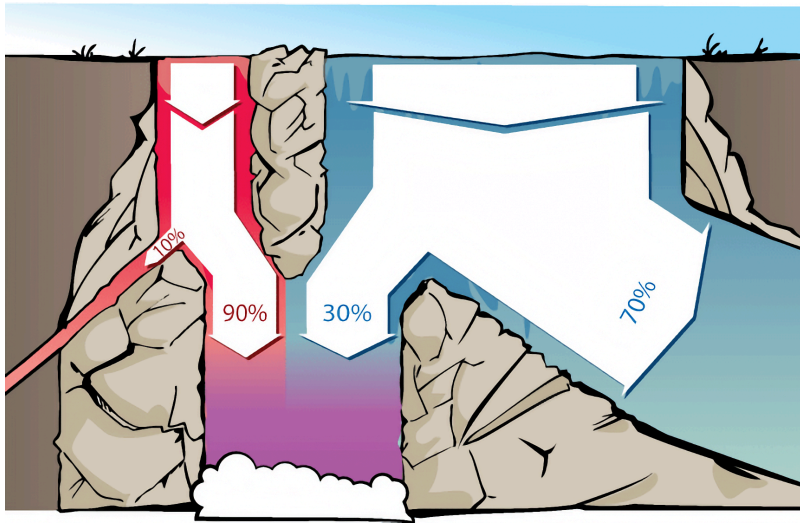
З тієї фіолетової води, що стекла в нижній басейн, яка частина налилася з червоного русла, а яка із синього?

1. 1 американський галон = 3,785411784 літри, тож 20 гал ≈ 75,71 л, 80 гал ≈ 302,83 л.

Ті 20% хворих пацієнтів аналогічні до 20 галонів червоної води за секунду, а 80% здорових — до 80 галонів синьої води:



Ті 90% хворих пацієнтів, у контакті з якими лопатка для язика чорніє, аналогічні до 90% червоної води, що стікає до підніжжя водоспаду. 30% здорових пацієнтів із позитивним результатом індикатора, аналогічні до 30% синьої води, що стікає до нижнього басейну.



Отже, питання «Яка частина води натекла у нижній басейні з червоного русла?» має таку ж відповідь, що й питання «Яка частина пацієнтів, хворих на дисизит, отримує позитивний результат?».

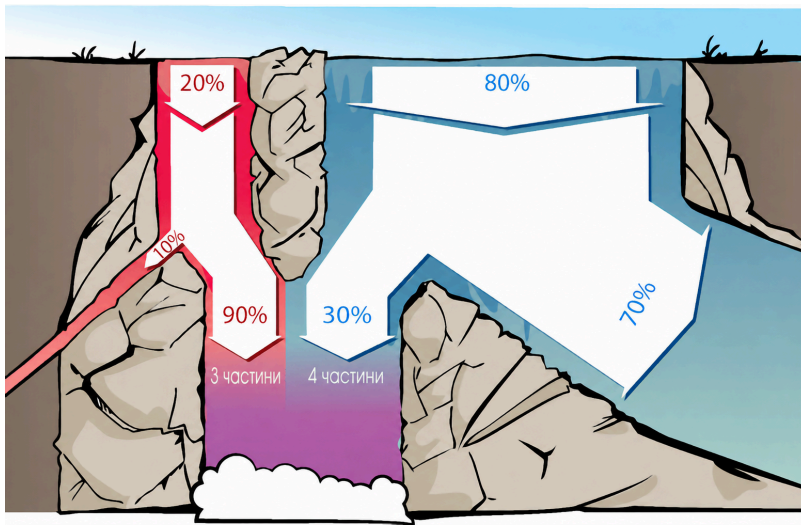
Почнемо з того, що на уступі водоспаду *синьої води вчетверо більше, ніж червоної*.

Уже тоді кожна молекула червоної води з імовірністю 90% потрапить у нижній басейн, а кожна молекула синьої води потрапить у басейн з імовірністю 30%. (А 90% червоної і 30% синьої води стече до дна.) Отже, ймовірність того, що кожна молекула червоної води потрапить на дно, *втричі більша* ($0{,}9/0{,}3 = 3$), порівняно з ймовірністю для молекули синьої води.

[Завдання цього есею — пояснити ймовірності, тому вживання слова «ймовірно» — у деякому сенсі приклад циклічного мислення. А зрештою, що означає «ймовірно»? Тут ми не вдаємося до таких подробиць. Імовірність передбачає певну частоту результатів, а тому твердження вище — спроба визначити частоту результатів (щодо пропорцій червоної та синьої води, що

проходить через водоспадну систему) шляхом посилання на іншу частоту результатів; звідси й виникає циклічність. Краще триматися саме пропорцій: скільки б не вилилося червоної води, стіче втричі більше ніж синьої. Скажімо, певна частка (f) молекул синьої води стікає вниз; тоді на кожні 100 молекул води проходить $f \times 80$ синіх молекул і $3f \times 20$ червоних, що дає співвідношення 60f червоних до 80f синіх. Ці пропорції можна зменшити діленням обох чисельників на f — отримуємо 60: 80, тобто 3: 4.]

Тож ми множимо апріорні пропорції 1: 4 (червоної води до синьої) на відносні правдоподібності 3: 1 і отримуємо кінцеві — $(1 \times 3): (4 \times 1) = 3: 4$, тобто у нижньому басейні до 3 частин червоної води припадає 4 частини синьої.



Щоб перевести ці *відносні* пропорції в *абсолютну* ймовірність того, що випадкова молекула води на дні червона, ми рахуємо $3/(3 + 4)$, щоб побачити, що 3/7 (близько 43%) води у нижній басейн витекла з червоного русла.

Ця пропорція не відрізняється від пропорції 18: 24 хворих пацієнтів із позитивними результатами до здорових пацієнтів із позитивними результатами, яку ми отримали б, якби думали в рамках ста пацієнтів.

Тобто, щоб у голові розв'язати задачу із дисизітом, ви можете перетворити цю текстову задачу:

> 20% пацієнтів у скринінговій популяції хворіють на дисизіт. У 90% пацієнтів з дисизітом і в 30% пацієнтів без нього чорніє лопатка для язика. Яка ймовірність того, що у пацієнта дисизіт, за умови, що його лопатка для язика почорніла?

До розрахунків:

> Такс, початкові шанси — (20%: 80%) = (1: 4), правдоподібності — (90%: 30%) = (3: 1). Перемноживши ці співвідношення, отримуємо шанси (3: 4), тобто ймовірність 3/7.

Як перемножити шанси, якщо ви не знайомі з математикою

> Спочатку у нас є шанси 20%: 80%. Для такого запису шансу можна прибрати знак відсотка. Поділивши обидві частини на найменше число (20), отримаєте охайний і простий запис — 1: 4. Він читається як 1 «шанс» для червоного і 4 «шанси» для синього. Правдоподібність того, що вода сягне дна, становить 90% для червоного русла і 30% для синього, тобто можна записати як 90: 30. Якщо ви поділите обидві сторони на 30, знову отримаєте гарний запис — 3: 1. Трійка позначає шанс того, що червоне русло стіче до дна, одиниця — що синє.
>
> Щоб об'єднати їх, помножте два числа, що позначають червоне, і залиште їх ліворуч. Те саме зробіть для синього і залиште їх праворуч. 1 (початковий) помножити на 3 (правдоподібність) = 3 (кінцевий) для червоного, а 4

(початковий) помножити на 1 (правдоподібність) = 4 (кінцевий) для синього, що дає кінцевий шанс 3:4. Щоб перетворити це число назад у відсотки, згадайте, що червоне має 3 шанси порівняно з 4 для синього. Між ними усього 3 + 4 = 7 варіантів. У червоного — 3 з них, тому шанси 3/7 або 0,43. Як і годиться для відсотків, просто помножте десятковий дріб на 100 і матимете 43%.

(Можливо, у вас і не вийде в голові перетворити 3/7 на 43%, але на око прикиньте, що ця частина помітно менша за 50%.)

Для цієї задачі можете спробувати схожі обрахунки:

- 90% віджетів робочі, 10% несправні.
- 12% робочих віджетів іскряться.
- І лише 4% робочих мають ту саму рису.

Який відсоток віджетів, що іскряться, несправний? Якщо вам зрозуміла умова задачі, спробуйте розв'язати її в голові від початку до кінця.

(Можете спробувати уявити водоспад із робочими та неробочими віджетами на уступі, з яких лише ті, що іскряться, спадають у нижній басейн.)

Відповідь: Маємо 1:9 несправних віджетів до робочих. Відносна правдоподібність того, що несправні віджети заіскряться відносно справних, становить 12:4. Спрощуємо до $(1:9)x(3:1) = (3:9) = (1:3)$, тобто 1 неробочий іскристий віджет до 3 робочих іскристих. Перетворюємо на ймовірності: $1/(1 + 3) = 1/4 = 25$ — тобто 25% іскристих віджетів несправні. Те, що ми бачимо іскри, не дає нам підстав вірити, що «віджет несправний», адже ймовірність сягнула лише 25%, менше ніж 50/50. Але з цього не випливає, що ми досі віримо, що віджет робочий, і можемо просто відкинути свідчення чи нехтувати ними. Несправний віджет *відносно ймовірніше* буде іскритися, а тому, отримавши таке свідчення, ми маємо вірити, що

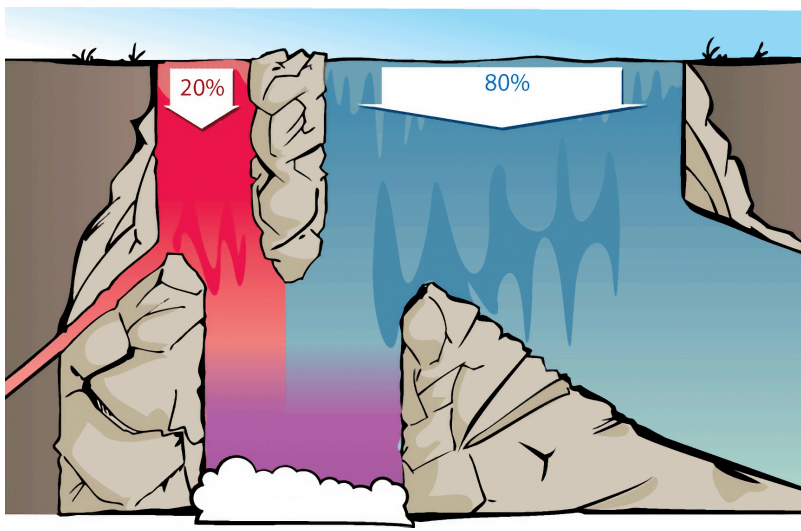
ймовірність того, що пристрій неробочий, *відносно вища*, навіть якщо вона не стала більшою за 50%. Ми підвищуємо нашу оцінку ймовірності з 10% до 25%.

Вступ до теореми Баєса: форма співвідношення шансів

У загальному вигляді теорема Баєса твердить:

Апріорні шанси × Відносні правдоподібності = Апостеріорні шанси

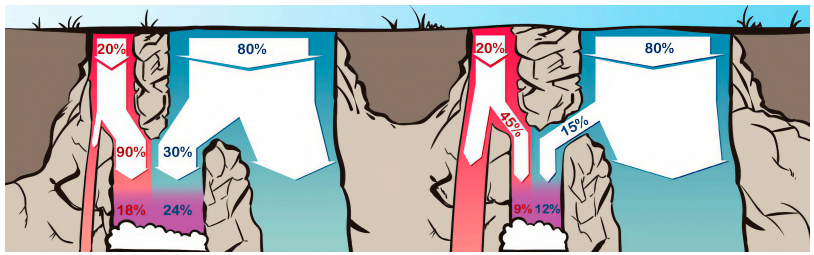
Ознайомившись із каскадною візуалізацією або прикладом із дисизітом, ми можемо уявити, наскільки *відносні шанси* — відповідний спосіб міркувати про задачу із двома руслами на вершині водоспаду.



Пропорція червоної до синьої води у нижньому басейні буде однаковою, хоч водоспадом стікає 200 галонів червоної води до 800 галонів синьої за секунду, 20 000 до 80 000 гал/с чи 1 до 4 гал/с. Поки решта водоспаду поводиться пропорційно, ми будемо отримувати ту саму пропорцію червоної води до синьої у

нижньому басейні. Отже, ми можемо нехтувати *обсягами* води та враховувати лише відносну пропорцію між обсягами.

Так само має значення *відносна* пропорція між тим, скільки з кожного русла потрапляє червоної і синьої води у спільний басейн. 45% червоної і 15% синьої води, що стікає до нижнього басейну, дадуть ту саму *відносну пропорцію*, що й 90% червоної і 30% синьої води у спільному басейні.



Такий підхід дозволяє нехтувати конкретними даними — 90% води з червоного русла і 30% із синього стікають до нижнього басейну — і звести їх до *відносних правдоподібностей* (3: 1).

Більш загальний випадок: скажімо, є медичний тест, що виявляє захворювання з частотою істинно позитивних результатів 90% (відповідно, хибно негативних — 10%) і частотою хибно позитивних результатів 30% (відповідно, істинно негативних — 70%). Позитивний результат цього тесту має таку ж *силу свідчень*, що й тест, де 60% істинно позитивних і 20% хибно позитивних результатів. Негативний же результат має ту ж *силу свідчень*, що й тест із 9% хибно негативних результатів і 63% істинно негативних.

Загалом, сила свідчень визначається тим, наскільки різні стани світу роблять наше спостереження ймовірним *порівняно* один з одним.

Рівняння

Щоб сформулювати теорему Баєса у повній загальності та довести її як теорему, нам потрібно ввести деякі нові позначення.

Умовна ймовірність

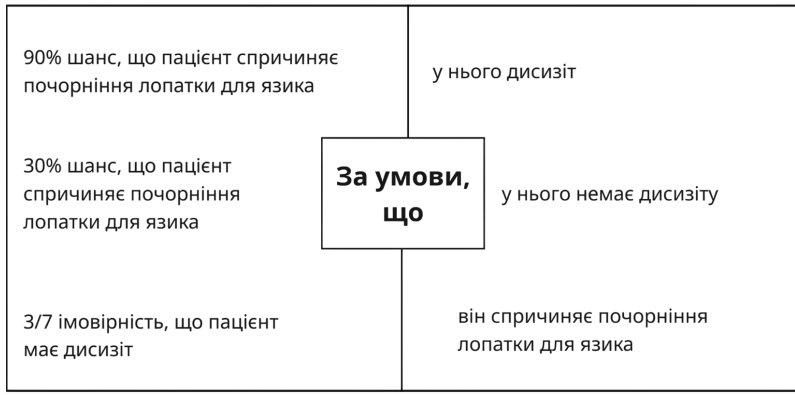
Якщо X є твердженням, то $P(X)$ — ймовірність X.

Іншими словами, X — твердження, що насправді або істинне, або хибне, про що ми не певні. $P(X)$ — спосіб виразити ступінь переконаності в тому, що X істинне. На ділі ж пацієнт або хворий, або здоровий, але якщо ви не знаєте його стану, свідчення можуть привести вас до 43% суб'єктивної ймовірності, що пацієнт хворий.

$\neg X$ означає, що «X хибне», тож $P(\neg X)$ — «ймовірність, що X хибне».

Проте приклад із дисизітом передбачав дещо складніші твердження, зокрема:

- Шанс, що пацієнт спричиняє почорніння лопатки для язика *за умови*, що у нього дисизіт, становить 90%.
- Шанс, що пацієнт спричиняє почорніння лопатки для язика *за умови*, що у пацієнта немає дисизіту, становить 30%.
- Імовірність, що пацієнт хворий на дисизіт *за умови*, що він спричинив почорніння лопатки, становить 3/7.



У таких випадках ми хочемо перейти від певного твердження, що *вважається* або *визнається* істинним (праворуч), до іншого твердження (ліворуч), нову ймовірність якого ми хочемо дізнатися, з урахуванням цього припущення.

Ось такі ймовірнісні твердження відомі як «умовні ймовірності». У стандартному записі умовної ймовірності наведені вище величини позначаються так:

- $P(\text{почорніла}|\text{хворий}) = 0{,}9$
- $P(\text{почорніла}|\neg\text{хворий}) = 0{,}3$
- $P(\text{хворий}|\text{почорніла}) = 3/7$

Цей стандартний запис для $P(X \mid Y)$ — ймовірність X за умови, що Y істинне, — на жаль, є лише симетричною вертикальною лінією, що не дає жодної візуальної підказки для запам'ятовування того, що припущення знаходиться праворуч, а виведене твердження — ліворуч. Умовна ймовірність визначається (запис «:=» означає «визначається як») таким чином (запис $X \wedge Y$ позначає «X та Y» або «і X, і Y істинні»):

$$P(X \mid Y) := \frac{P(X \wedge Y)}{P(Y)}$$

У прикладі з дисизітом P(хворий | почорніла) обчислюється діленням 18% — частки учнів, які одночасно хворі й мають почорнілу лопатку (P(хворий ∧ почорніла)), — на загальні 42% учнів, у яких лопатка просто почорніла (P(почорніла)).

Або P(почорніла|¬хворий) — імовірність почорніння лопатки *за умови*, що учень здоровий, — обчислюється як 24 (учні, які одночасно здорові *та* мають почорнілу лопатку), поділене на 80 (здорові учні). $24/80 = 3/10$, що відповідає 30% хибно позитивних результатів, про які нам говорили на початку.

Закон умовної ймовірності можна розглядати так: «Розглядатимемо лише ті світи або об'єкти, для яких Y істинне. Усередині цієї множини визначимо, скільки випадків одночасно задовольняють X та Y».

Теорема Баєса

Теорема Баєса твердить:

Апріорні шанси × Відносні правдоподібності = Апостеріорні шанси

У прикладі з дисизітом вона б формулювалася так:

$$\frac{P(\text{хворий})}{P(\text{здоровий})} \times \frac{P(\text{почорніла} \mid \text{хворий})}{P(\text{почорніла} \mid \text{здоровий})} = \frac{P(\text{хворий} \mid \text{почорн}}{P(\text{здоровий} \mid \text{почор}}$$

Апріорні шанси стосуються відносної пропорції хворих до здорових пацієнтів — 1:4. Перевівши відношення шансів на ймовірності, отримаємо $P(\text{хворий}) = \frac{1}{4+1} = \frac{1}{5} = 20\%$.

Відносна правдоподібність стосується того, наскільки ймовірно, що позитивний результат отримають хворий і здоровий пацієнт. Скористаємося позначеннями умовних імовірностей: $\frac{P(\text{позитивний} \mid \text{хворий})}{P(\text{позитивний} \mid \text{здоровий})} = \frac{0{,}90}{0{,}30}$, тобто відносні правдоподібності 3:1.

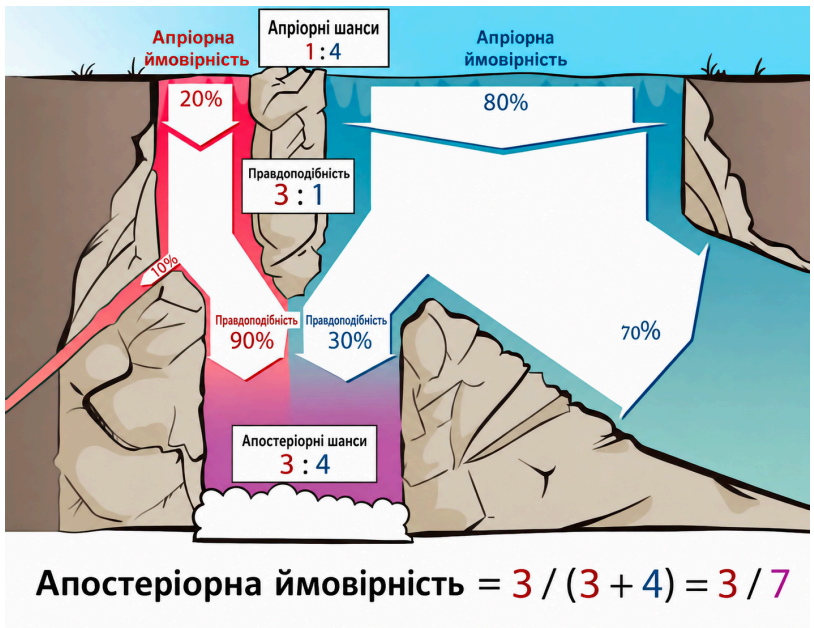
Апостеріорні шанси — відносні пропорції хворих до здорових пацієнтів серед тих, хто отримав позитивні результати тесту, або $\frac{P(\text{хворий} \mid \text{позитивний})}{P(\text{здоровий} \mid \text{позитивний})} = \frac{3}{4}$, тобто шанси 3: 4.

Щоб дістати *ймовірності* з відносних шансів, потрібно пам'ятати, що ймовірності взаємно виключних і вичерпних тверджень мусять дорівнювати одиниці, тобто 100% того, що щось станеться. Оскільки усі або не хворі, або хворі, можна нормалізувати співвідношення шансів 3: 4 через ділення на суму членів:

$(\frac{3}{3+4}: \frac{4}{3+4}) = (\frac{3}{7}: \frac{4}{7}) \approx (0{,}43: 0{,}57)$

... що дає нам ймовірності (0,43: 0,57), пропорційні до початкового співвідношення (3: 4), але він уже дорівнює одиниці. Бо якби ймовірність настання якоїсь події дорівнювала 3 (300%), то було б щось геть дивне.

Уявити це нам допоможе каскадна візуалізація:



Усі ці розрахунки можна узагальнити до *будь-якої* з двох **гіпотез** H_j та H_k зі **свідченням** e, у випадку чого теорему Баєса можна розписати так:

$$\frac{P(H_j)}{P(H_k)} \times \frac{P(e \mid H_j)}{P(e \mid H_k)} = \frac{P(H_j \mid e)}{P(H_k \mid e)}$$

(читається як «апостеріорне відношення шансів для гіпотез H_j проти H_k (після спостереження свідчення e) дорівнює апріорному відношенню шансів, помноженому на відношення того, наскільки добре гіпотеза H_j передбачила це свідчення порівняно з гіпотезою H_k»).

Якщо H_j та H_k взаємно виключні та вичерпні, можна перетворити ці апостеріорні шанси в апостеріорну ймовірність для H_j шляхом нормалізації шансів: поділити співвідношення шансів на суму його членів, щоб сума складників нового співвідношення дорівнювала 1.

Доведення теореми Баєса

Переформулюємо визначення умовної ймовірності: $P(X \wedge Y) = P(Y) \times P(X|Y)$. Наприклад, щоб встановити «частку всіх хворих пацієнтів, які отримують позитивний результат», множимо «частку хворих пацієнтів» на «ймовірність, що у хворого пацієнта почорніє лопатка для язика».

Так і отримуємо доведення теореми Баєса, де e_0 — нове свідчення:

$$\frac{P(H_j)}{P(H_k)} \cdot \frac{P(e_0 \mid H_j)}{P(e_0 \mid H_k)} = \frac{P(e_0 \wedge H_j)}{P(e_0 \wedge H_k)} = \frac{P(H_j \mid e_0)/P(e_0)}{P(H_k \mid e_0)/P(e_0)} = \frac{P(H_j \mid e_0)}{P(H_k \mid e_0)}$$

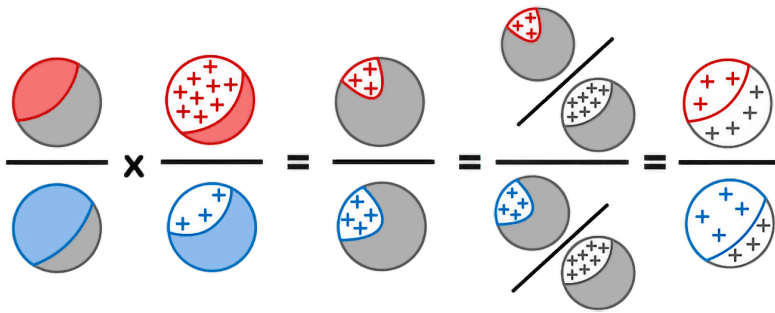
що й треба було довести.

У прикладі із дисизітом ці етапи доведення відповідають таким операціям: було б корисно, якби ви пригадали ймовірності

з прикладу про дисизіт, перед тим, як я покажу формулу.

$$\frac{0{,}20}{0{,}80} \times \frac{0{,}90}{0{,}30} = \frac{0{,}18}{0{,}24} = \frac{0{,}18/0{,}42}{0{,}24/0{,}42} = \frac{0{,}43}{0{,}57}$$

На візуалізації етапів обрахунку червоним кольором позначені хворі, синім — здорові, сірим — поєднання хворих і здорових пацієнтів, плюсом позначені позитивні результати тестів:



Цей процес спостереження за свідченнями і застосування їхнього співвідношення правдоподібності для перетворення апріорного переконання на апостеріорне зветься «баєсовим оновленням» або «переглядом переконання».

- Про узагальнення теореми Баєса у формі співвідношення шансів до багатьох гіпотез і свідчень читайте у «Теорема Баєса: векторна форма».
- Про трансформацію форми шансів, що ще більше унаочнює силу свідчень, читайте «Теорема Баєса: логарифмічні відношення шансів».

Теорема Баєса: пропорційна форма

Якщо H_i та H_j — гіпотези, а e — свідчення, тоді теорема Баєса твердить:

$$\frac{P(H_i)}{P(H_j)} \times \frac{P(e \mid H_i)}{P(e \mid H_j)} = \frac{P(H_i \mid e)}{P(H_j \mid e)}$$

У задачі із дисизітом хворобу мають 20% пацієнтів з цільової популяції, у 90% хворих і 30% здорових пацієнтів спостерігаємо почорніння індикаторної смужки. Цю задачу можна розв'язати за допомогою формули Баєса, наведеної вище: $(1:4) \times (3:1) = 3:4$.

Якщо замість співвідношення шансів обчислити числа для кожного члена рівняння, натомість отримаємо такі розрахунки: $\frac{1}{4} \times \frac{3}{1} = \frac{3}{4}$ або $0{,}25 \times 3 = 0{,}75$.

Безпосередньо витлумачити цей запис можна так: «Якщо спочатку шанси того, що пацієнт хворий, проти того, що він здоровий, становлять 0,25, а результат аналізу втричі правдоподібніший за умови, що пацієнт хворий, ніж за умови, що він здоровий, то робимо висновок, що шанси того, що пацієнт хворий, проти того, що він здоровий, становлять 0,75».

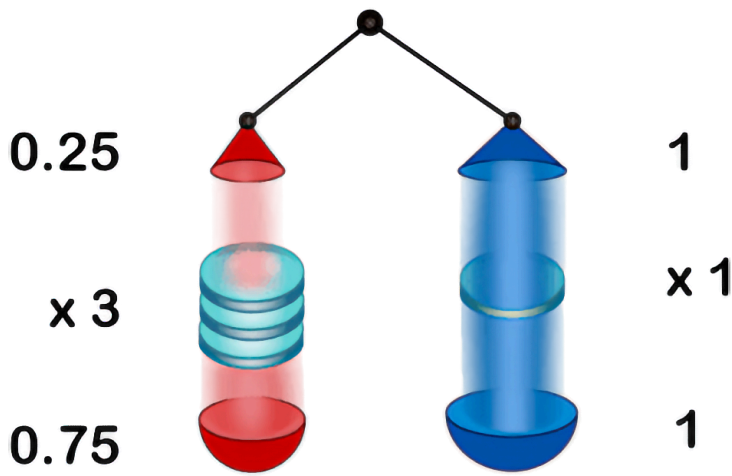
Так виглядає правильне міркування, і таку форму Баєсової теореми називаємо *пропорційною*. Якщо ми захочемо повернутися до ймовірнісної форми, тоді міркуємо так: якщо на кожного 1 здорового пацієнта припадає 0,75 хворих, маємо $0{,}75/(0{,}75 + 1) = 3/7 = 43\%$ хворих пацієнтів.

Прожекторна візуалізація

Один зі способів подивитися на цей результат: оскільки при множенні на додатну константу співвідношення шансів

еквівалентні, ми можемо зафіксувати праву сторону співвідношення як рівну одиниці та з'ясувати, що міститься зліва. Саме так ми і робимо, коли бачимо розрахунок типу $(0{,}25:1)\cdot(3:1)=0{,}75:1$ — форму, запропоновану доведеною вище теоремою.

Ми можемо візуалізувати теорему Баєса як пару прожекторів із різною початковою інтенсивністю, світло від яких проходить крізь лінзи, що посилюють або послаблюють кожну вхідну одиницю світла на фіксований множник. У випадку із дисизітом, якщо ми зафіксуємо для правого синього променя початкову інтенсивність 1 і множник лінзи 1, а для лівого променя — початкову інтенсивність 0,25 і множник лінзи 3x, то отримаємо візуалізацію обчислення, передбаченого теоремою Баєса:



Зауважте подібність прожекторної візуалізації до каскадної. Головне, що додає прожекторна візуалізація, — вона допомагає уявити абсолютну інтенсивність джерел світла та лінз і при

цьому зберегти відносну так, щоб промені й лінзи праворуч мали значення, рівні одиниці.

Застосування для побудови аргументу

Певно, пропорційна форма Баєсової теореми — найшвидший спосіб описати баєсове міркування, і звучить вона так, ніби має бути істинною. Якби ви прописували художнього персонажа, який зненацька видає баєсівський аргумент посеред історії, а її читачі якої ні сном ні духом про Баєсову теорему, мабуть, ви б вклали у його вуста такі слова:

> «Припустимо, що мітка точно продовжить своє існування, якщо розум Темного Лорда живе, але заздалегідь ми б припустили, що з імовірністю 20 відсотків Чорна мітка продовжить існування після смерти Темного Лорда. Тоді спостереження "Чорна мітка не зникла" вп'ятеро імовірніша у світах, де Темний Лорд живий, ніж у світах, де він мертвий. Невже це можна порівняти з апріорною імовірністю безсмертя? Скажімо, апріорна оцінка того, що Темний Лорд виживе, — один до ста. Якщо гіпотеза в сто разів імовірніше хибна, ніж істинна, а тоді ми бачимо свідчення вп'ятеро разів імовірніше за умови, що ця гіпотеза істинна, треба оновити віру в гіпотезу до «у двадцять разів більш імовірно хибна, ніж істинна». Шанси один до ста, помножені на співвідношення імовірності один до п'яти дорівнює шансам один до двадцяти, що Темний Лорд мертвий...».[1]

1. Узяв з українськомовного перекладу «Гаррі Поттер і Методи Раціональности» (Розділ 86. Перевірка чисельних гіпотез (https://xn--c1asif2i.xn--j1amh/086.html)). Невеличке зауваження: у цьому есеї йдеться не стільки про ймовірності, скільки про шанси — співвідношення двох взаємовиключних альтернатив (шанс того, що пацієнт «хворий:здоровий », що гіпотеза «істинна:хибна », що Темний Лорд « живий:мертвий »). — Прим. пер.

Схожим чином, якби ви як лікар намагалися пояснити пацієнту, що означає позитивний результат аналізу, напевно, сказали б таке: «Якби ми ще не бачили результатів аналізу, шанси на те, що такі пацієнти, як ви, здорові, проти шансів на те, що вони хворі, становили б 1000: 1. Цей тест лише у 100 разів правдоподібніше дає позитивний результат у хворих пацієнтів, ніж у здорових. Тож тепер ми вважаємо, що шанси на те, що ви здорові, проти шансів на те, що ви хворі, становлять 10: 1 — і це все ще досить непогані шанси!».

Завдяки наочним діаграмам і спеціальним записам для шансів та відносних правдоподібностей Баєсова теорема може виглядати більш інтуїтивною, але пропорційна форма — чи не найкращий приклад того, що *звучить* правдоподібно (але водночас *кількісно правильно*) і що можна ословити трьома реченнями.

Надзвичайні твердження потребують надзвичайно сильних свідчень

Баєсова теорема повідомляє нам, наскільки сильним має бути свідчення, щоб підтримати дану гіпотезу. Завдяки такій можливості ми можемо бачити, чи свідчення достатнє чи недостатнє, щоб підвищити ймовірність понад 50%.

Як приклад розглянемо задачу з іскристими віджетами:

> 10% віджетів неробочі, 90% — робочі. 4% робочих і 12% неробочих іскряться. Можете в голові порахувати, який відсоток іскристих віджетів не працює?

Апріорні шанси «віджет несправний» проти «віджет справний» становлять 1:9.

12% неробочих і 4% робочих віджетів видають іскри, тож співвідношення правдоподібності спостереження іскри за двох гіпотез становить 3:1 (неробочі віджети втричі ймовірніше видаватимуть іскри).

$(1:9) \times (3:1) = 3:9 \cong 1:3$ — апостеріорні шанси того, що іскристий віджет несправний, проти того, що він справний. Отже, 1/4 іскристих віджетів неробочі.

Апріорні шанси того, що віджет несправний, відносно малі — 1 з 10. Ми провели перевірку — шукали іскри — і дізналися, що тест був утричі ймовірніший за несправного віджета, ніж за справного. Тобто *це свідчення виявилося недостатньо сильним, щоб подолати апріорну неправдоподібність* твердження.

Це не означає, що ми викидаємо свідчення чи нехтуємо ними. Це означає, що після оновлення оцінки на підставі спостереження іскріння ми оцінили апостеріорну ймовірність як 25% того, що віджет несправний — вона не перетнула поріг 50%.

Що повинно змінитися, щоб посунути ймовірність того, що віджет несправний, вище 50%? Нам знадобиться свідчення з більш радикальним співвідношенням правдоподібності — радикальнішим за апріорні шанси 9: 1 проти гіпотези. Наприклад, якби несправні віджети іскрилися з імовірністю 50%, а справні віджети — з імовірністю 5%, апостеріорні шанси посунулися б до шансів 10: 9 або ймовірності 53%.

Іншими словами, щоб апріорно неправдоподібне твердження *X* отримало високу апостеріорну ймовірність, співвідношення правдоподібності для нового свідчення на підтримку *X* з-поміж альтернатив має бути *більшим* за силу апріорних шансів проти *X*.

Ось такий кількісний аргумент стоїть за якісним твердженням «надзвичайні твердження потребують надзвичайно сильних свідчень» (популяризованим Карлом Саганом, а вислів сягає щонайменше самого П'єра-Симóна Лапла́са)[1].

Тобто «надзвичайне твердження» — це те, у якого перед розглядом свідчень низька апріорна ймовірність, а «надзвичайно сильні свідчення» — ті, у яких надзвичайно велике співвідношення правдоподібності на підтримку цього твердження з-поміж альтернативних.

Що визначає надзвичайну силу свідчення?

Співвідношення правдоподібності визначається так:

$$\text{СП} = \frac{P(\text{свідчення} \mid \text{гіпотеза істинна})}{P(\text{свідчення} \mid \text{гіпотеза хибна})}$$

1. Ознайомитися детальніше із поняттям надзвичайності твердження можна у статті (https://defibering.wordpress.com/2023/01/20/scientific-method/#_8._%D0%A1%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82_%D0%A1%D0%B0%D0%B3%D0%B0%D0%BD%D0%B0) Андрія Шовкошитного «Науковий метод». — Прим. пер.

Щоб отримати надзвичайно велике співвідношення правдоподібності, знаменник дробу має бути дуже малим. А дуже великий чисельник сам по собі мало допомагає отримати екстремальне співвідношення. Навіть збільшення чисельника з 99% до 99,9999% майже нічого не змінить, якщо знаменник залишається 70%.

Отже, щоб отримати надзвичайно сильні свідчення, нам треба отримати спостереження, яке вкрай *малоймовірне* за умови звичного стану справ, але, згідно з надзвичайним твердженням, доволі ймовірне. Таке спостереження заслуговує на титул «надзвичайно сильне свідчення».

Приклад надзвичайного твердження і звичайного свідчення: прибульці з книжкових шаф

От така гіпотеза: а що як існують прибульці, які вночі телепортуються у ваші будинки і доставляють туди книжкові шафи?

Боб запропонував наступне свідчення для цього твердження: «Минулого тижня я ходив у гості до друга, він якраз купив собі нову книжкову шафу. Якби не існувало ніяких книгорозкидувальних прибульців, я б не очікував, що у друга з'явиться нова шафа. Але якщо такі прибульці існують, ймовірність того, що у друга з'являється нова книжкова шафа, значно вища. Отже, моє спостереження "Вдома у мого друга з'явилася нова книжкова шафа" дійсно слугує свідченням на підтримку існування прибульців із книжкових шаф».

Інтуїтивно ми розуміємо, що Бобове свідчення «Вдома у мого друга з'явилася нова книжкова шафа» не настільки надзвичайно сильне, як твердження «Існують прибульці, які живуть у книжкових шафах», що це свідчення не підвищує ймовірність настільки, щоб зробити гіпотезу переконливою. І Баєсова теорема формалізує це інтуїтивне твердження.

По суті, Боб правий: його спостереження «Вдома у мого друга з'явилася нова книжкова шафа» підтримує існування тих прибульців. Залежно від того, як давно Боб ходив у гості до

свого друга, за інших рівних умов може бути шанс 1% того, що у друга з'явилася нова шафа з книжками. З іншого боку, гіпотезі про прибульців із книжкових шаф можна призначити ймовірність 50%, що ті прибульці оберуть саме його будинок з-поміж інших. Якщо так, то співвідношення правдоподібності на підтримку гіпотези становить 50: 1.

Проте адекватна апріорна оцінка щодо існування прибульців з книжкових шаф дасть вкрай *низьку* апріорну ймовірність за умови наших інших попередніх спостережень про світ. Можемо дати стриману оцінку і шанси 1: 1 000 000 000 проти існування прибульців із книжкових шаф. Тоді, щоб підняти апостеріорну ймовірність нашої віри у прибульців з книжкових шаф до якогось «діапазону із практично помітною різницею» 1: 100, нам потрібно буде побачити свідчення з накопичувальним співвідношенням правдоподібності 10 000 000: 1 на підтримку прибульців з книжкових шаф. 50: 1 не вистачить.

Що повинно змінитися, щоб спостереження «Вдома у мого друга з'явилася нова книжкова шафа» стало переконливим свідченням на підтримку існування прибульців з книжкових шаф порівняно з нульовою гіпотезою «звичний стан справ»?

Згідно з баєсівською інтерпретацією сили свідчення, нам потрібно побачити щось таке, що практично неможливо, якщо прибульців у книжкових шафах *не* існує. Нам довелося б повірити, що, за умови істинності звичайного стану справ, правдоподібність побачити книжкову шафу становить близько 0,00000001%. У такому разі співвідношення правдоподібності — так звана вагомість свідчення — становило б приблизно мільярд до одного на користь прибульців порівняно зі звичайним станом речей.

І все одно варто ще подумати, чи існують інші альтернативні гіпотези, крім «прибульці, які живуть у книжкових шафах» і «звичайний стан справ»: наприклад, спланована кимось змова підкидати книжкові шафи у будинки. Але тоді хоч не маємо справу зі спостереженням, настільки очікуваним за нульової

гіпотези, що воно не може підкріпити *жодне* надзвичайне твердження.

Однак, якщо натомість ми припустимо, що ймовірність того, що ті прибульці встановлять книжкову шафу вдома у Бобового друга, становить буцімто 99,999999%, то ситуація майже не змінюється: відношення правдоподібності становить 99,999999%: 1%, або 100: 1, замість 50: 1. Щоб отримати дуже велике співвідношення правдоподібності, насамперед нам потрібен дуже малий знаменник, а не великий чисельник. Іншими словами — «надзвичайно сильне свідчення».

Що визначає надзвичайність твердження?

Логічно наступним кроком запитати, за рахунок чого твердження стає «надзвичайним» чи «звичайним». Це питання торкається окремої глибокої теми, але як приклад можна навести таке твердження: через надходження вуглекислого газу в атмосферу Земля нагрівається.

Щоб оцінити звичайність чи надзвичайність цього твердження:

- Ми *не* питаємо, чи майбутні наслідки цього твердження здаються надзвичайними чи важливими.
- Ми *не* питаємо, чи заходи, необхідні для реагування на твердження або його наслідки, дуже дорогі.
- Ми питаємо, чи «діоксид вуглецю нагріває атмосферу» чи «діоксид вуглецю не нагріває атмосферу» краще узгоджується з *глибокими причиновими* узагальненнями, які ми вже вивели про діоксид вуглецю і тепло.
- *Якщо* ми вже розглянули такі глибокі причинові узагальнення, то не питаємо про узагальнення, що *причиново випливають* із тих глибоких причинових узагальнень, які ми вже розглянули. (Наприклад, ми не кажемо: «Але протягом останніх двохсот років, упродовж яких проводилися спостереження, глобальна температура не

виходила за рамки нижченаведеного діапазону; а вихід за межі цього діапазону був би "надзвичайним"»).

Ці критерії показують, що твердження «Великі обсяги доданого діоксиду вуглецю поступово нагрівають атмосферу Землі» було б «звичайним» ще до початку пошуку будь-яких свідчень на користь чи проти твердження — адже саме так, більш-менш, і має поводитися парниковий газ. Отже, немає епістемічних підстав вимагати від цієї гіпотези, щоб вона давала прогноз, що за будь-якої іншої гіпотези був би вкрай малоймовірним, перш ніж у неї повірити.

Накопичувальне оновлення

Чільна особливість прикладу з прибульцями полягає в тому, що баєсівці визнають: спостереження нової книжкової шафи — локально одне свідчення зі співвідношенням правдоподібності 50: 1 на підтримку існування прибульців із книжкових шаф. Баєсівець не вишвирює спостереження з вікна через його *недостатність* — натомість вони накопичуються в сукупності свідчень. Якщо ви ходите від будинку до будинку і бачите нову книгошафу за новою книгошафою, баєсіанець волі-поволі починає задумуватися, чи не відбувається щось дивне, а не відкидає кожне спостереження як «недостатнє свідчення», а потім забуває про них.

Такий підхід контрастує зі способом, у який люди часто поводяться інстинктивною: дійшовши висновку, що на підставі побачених свідчень не слід вірити у прибульців з книжкових шаф, вони повністю відкидають те свідчення, зрікаються його і твердять, що свідченням воно ніколи й не було. (Це прояв ставлення до аргументів як до солдатів: будь-яке свідчення на користь якогось твердження має бути «переможене».)

Баєсівець же просто скаже: «Так, це свідчення підтримує це твердження, але воно кількісно недостатнє». Цей вираз теж контрастує з практикою ставитися до будь-яких поступок, на які йде опонент, як до перемоги. Якщо вважати, що всі аргументи на підтримку істинних тверджень мають підтверджуватися, а всі

аргументи противника — спростовуватися, то навіть один-єдиний непереможний аргумент на *підтримку* твердження слугує доказом її перемоги, незалежно від того, наскільки сильними чи слабкими є свідчення, які він надає. Але не для баєсівців — вони вважають, що спостереження книжкової шафи може бути локальним свідченням на підтримку існування прибульців з книжкових шаф, просто цього свідчення дуже малувато, щоб повірити в тих прибульців.

Свідчення, що перекривають попередню оцінку

Якщо ви думаєте, що у якогось твердження апріорні шанси $1:10^{100}$ і вам хтось надає свідчення зі співвідношенням правдоподібності $10^{94}:1$ на підтримку цього твердження, не поспішайте казати: «О, гадаю, апостеріорні шанси становлять $1:10^{6}$ ». Натомість варто запитати себе, чи (а) ви не помилилися з апріорними шансами або (б) свідчення не настільки сильне, як ви його оцінили. Повірити у гіпотезу з дуже низькими апріорними шансами, зрештою, не так вже й складно. Наприклад, щоразу, коли ви дивитеся на послідовність з десяти цифр, підібрану генератором випадкових чисел, починаєте вірити в гіпотезу, у якої апріорні шанси близько десяти мільярдів до одного.

Але так має бути лише з істинними гіпотезами. А знайти сильну підтримку для хибних гіпотез можна набагато рідше. «Сильне свідчення» — це саме «такого роду свідчення, якого ми майже *ніколи* не бачимо, коли твердження виявляється хибним».

Уявіть: ви підкидаєте монетку не більше ніж 300 разів і питаєте, чи послідовність випадінь гербів і чисел дає більшу підтримку гіпотезі «на цій монеті герб випадає у 3/4 випадків», причому набагато сильнішу, ніж істинній гіпотезі «ця монета врівноважена». Як можна пересвідчитися цим кодом[2],

2. https://gist.github.com/Soares/941bdb13233fd0838f1882d148c9ac14

послідовність підкидань в якийсь момент підтримає хибну гіпотезу на рівні 10: 1 у 8% випадків, на рівні 100: 1 у 0,8% випадків і на рівні 1000: 1 у 0,08% випадків. (Зауважте: що сильніше нас вводить в оману спостереження, то менш імовірно, що така помилка трапиться.)

Ситуація, коли ви бачите свідчення вагою 10^{94} : 1 (94 порядків величини / 312 бітів), що підкріплюють хибну гіпотезу, має траплятися з вами, в середньому, лише раз на... ТАКОГО НІКОЛИ НЕ БУЛО.

Спостерігати подію, ймовірність якої справді становить 10^{-94} за умови хибності гіпотези, у випадку, коли гіпотеза *насправді хибна*, не доведеться нікому навіть разу протягом очікуваного часу існування Всесвіту.

Тому якщо ви думаєте, що апріорні шанси для монети становлять 1: 10^{100} проти неврівноваженості, а потім побачили 312 підкидань монети, з яких щоразу випадає герб... не подумайте: «Що ж, мої оновлені апостеріорні шанси проти неврівноваженості монети тепер ставлять 1: 10^{6} ». Тепер ви знаєте краще і подумаєте: «Гадаю, я помилявся, що апріорні шанси аж настільки низькі».

Звичайні твердження потребують свідчень звичайної сили

Якщо надзвичайні твердження вимагають надзвичайно сильних свідчень, тоді й звичайні твердження вимагають звичайних свідчень.[1]

Певно, цей логічний висновок відіграє важливішу роль на практиці саме тому, що у повсякденному житті ми частіше маємо справу зі звичайними твердженнями і рідше — з надзвичайними. Але можемо спокуситися тим, щоб відкидати численні звичайні твердження і вимагати надзвичайних свідчень (це упередження — відоме як «заангажований скептицизм»).

Приклад: новий працівник, який не справляється зі своїми обов'язками

Один із запропонованих прикладів:

> Кілька років тому моя старша колега казала, що новий працівник не виконував свою роботу вчасно, що їй доводилося контролювати кожен його крок, щоб він узагалі бодай щось зробив. З багатьох причин було неприємно чути таке. Мені подобався той хлопець — кількома тижнями раніше я виступав перед радою директорів за його найм (саме з цієї причини старша керівниця і звернулася до мене). Я міг би попросити більше свідчень, я міг би попросити більше часу, щоб ситуація владналася, я міг би попросити відеозапис та письмове затвердження... але те, що новачок не виправдовував очікувань, було *не таким уже й малоймовірним*. Чи міг я назвати точні апріорні шанси того, що новачок не виконуватиме своїх

1. Вислів про звичайні твердження і звичайні свідчення приписують Ґверну Бренвену (https://gwern.net/), який не популяризував його так, як Карл Саган популяризував свою версію. — Прим. авт.

обов'язків? Чи міг я сказати, наскільки ймовірніше було почути від менеджерки повідомлення про те, що працівник не справляється в довгостроковій перспективі, ніж похвалу працівнику, який у довгостроковій перспективі працює добре? Ні, але ситуація «хтось узяв не ту людину» трапляється постійно, і я вже бачив такі випадки. Твердження не так уже й сильно відхилялося від звичайного досвіду, і я не міг вимагати надзвичайно сильних свідчень. Мовою чисел, я думав, що пропорція «поганих до хороших працівників» становила близько 1:4 на той час, а правдоподібність повідомлення менеджерки за двох альтернативних гіпотез становила десь 10:1.

Іншими словами, правило звучить як «надзвичайні твердження потребують надзвичайно сильних свідчень», а не «незручні, але звичайні твердження потребують надзвичайно сильних свідчень».

У повсякденному житті ми маємо справу з більшою кількістю звичайних тверджень, ніж надзвичайних, включно з багатьма звичайними твердженнями, незручність яких може спокусити нас відкинути їхні адекватні, але свідчення звичайної сили. У таких випадках важливо пам'ятати: звичайні твердження потребують свідчень звичайної сили.

Теорема Баєса: логарифмічна форма шансів

Форма Баєсової теореми у вигляді шансів твердить, що апріорні шанси, помножені на співвідношення правдоподібності, дорівнюють апостеріорним шансам. Візьмемо логарифм з обох сторін рівняння і отримаємо еквівалентне рівняння, в якому члени додаються, а не множаться.

Нехай H_i та H_j позначають гіпотези, а e — свідчення, тоді логарифмічна форма Баєсової теореми у вигляді шансів виглядатиме так:

$$log(\frac{P(H_i \mid e)}{P(H_j \mid e)}) = log(\frac{P(H_i)}{P(H_j)}) + log(\frac{P(e \mid H_i)}{P(e \mid H_j)}).$$

Така форма може бути зручною, коли ви послідовно виконуєте багато оновлень. Але ще важливіший привід мислити у логарифмічних шансах — краще засвоїти «силу свідчення».

Логарифми співвідношень правдоподібності

Скажімо, ви плануєте навідати своїх друзів — подружжя Ендрю та Бетті. Вони пообіцяли, що після вашого прильоту хтось із них забере вас з аеропорту. Ви не певні, хто саме (апріорні шанси 50: 50), але знаєте ось що:

1. У них обох по машині — синя і червона. Ендрю надає перевагу синій, а Бетті — червоній, хоча кореляція відносно слабка. (Іноді їхній вибір залежить від того, на якій їздить їхня дитина.) Ендрю вдвічі частіше їздить на синій машині, ніж Бетті.
2. Як правило, Бетті сигналить, щоб її помітили. Ендрю теж сигналить, але рідше. Бетті вчетверо ймовірніше буде сигналити, ніж Ендрю.
3. Як правило, Ендрю запізнюється (частіше за Бетті). Імовірність того, що автомобіль Бетті вже буде в аеропорту

до вашого прильоту, вдвічі вища.

Усі три спостереження незалежні, наскільки вам відомо (тобто ви не припускаєте, що Бетті приїде раніше чи пізніше, якщо поїде на синій автівці тощо).

Скажімо, ми бачимо синю автівку: вона вже стоїть біля аеропорту і сигналить.

Розрахунок у формі шансів виглядатиме так: апріорні шанси 1:1 для «Бетті проти Ендрю» помножити на співвідношення правдоподібності $(1:2)\times(4:1)\times(2:1)$ дорівнює апостеріорним шансам $(1\times4\times2:2\times1\times1)=8:2=4:1$, тож імовірність 4/5 або 80%, що це Бетті.

Ось логарифмічна форма шансів тих самих розрахунків, де 1 **біт** позначає кожну двократну зміну у ступені переконаності або силі свідчення:

- Апріорна віра в Бетті $log_2(\frac{1}{1})=0$ бітів.
- Свідчення $log_2(\frac{1}{2})=-1$ біт проти Бетті.
- Свідчення $log_2(\frac{4}{1})=+2$ біти за Бетті.
- Свідчення $log_2(\frac{2}{1})=+1$ біт за Бетті.
- Апостеріорне переконання: $0+^-1+^+2+^+1=^+2$ біти за те, що нас підбере Бетті.

Якщо ваша апостеріорна віра дорівнює +2 біти, тоді апостеріорні шанси складають $2^{+2}:1=4:1$, що дає апостеріорну ймовірність 80%, що вас забере Бетті.

Свідчення і переконання, представлені у такому вигляді, просто *додаються одне до одного* (є адитивними), що краще підходить для інтуїтивного розуміння «сили переконаності» та «сили свідчення»; далі розглянемо цю думку детальніше.

Вісь логарифмічних шансів

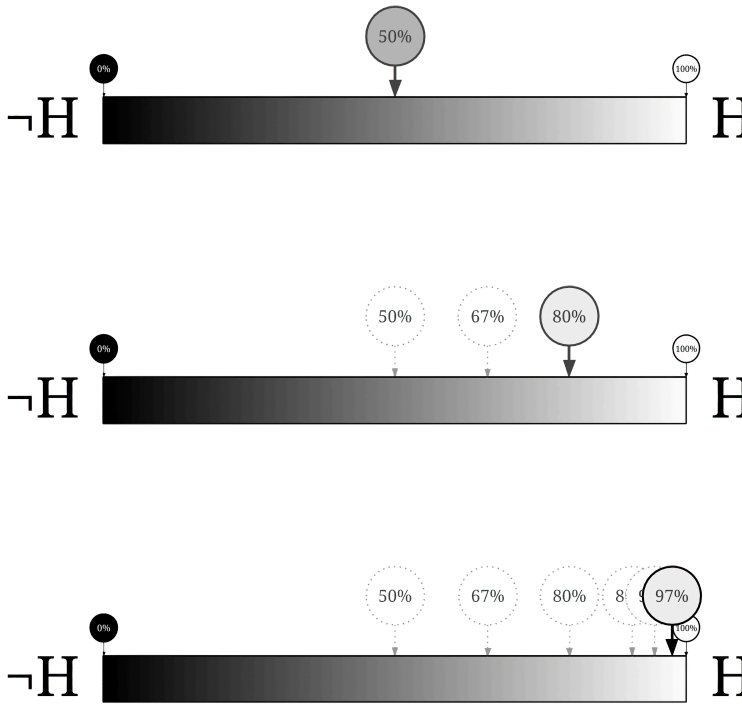
Нехай ви вважаєте, що гіпотеза H настільки ж імовірна, як і $\neg H$ (її заперечення). Тоді ви отримуєте п'ять окремих незалежних

оновлень 2:1 на користь *H*. Що стається з вашими ймовірностями?

Ваші шанси (для *H*) перетворюються таким ланцюжком: $1:1 \to 2:1 \to 4:1 \to 8:1 \to 16:1 \to 32:1$.

Тож ваші ймовірності змінюються так: $\frac{1}{2} = 50\% \to \frac{2}{3} \approx 67\% \to \frac{4}{5} = 80\% \to \frac{8}{9} \approx 89\% \to \frac{16}{17} \approx 94\% \to \frac{32}{33} \approx 97\%$.

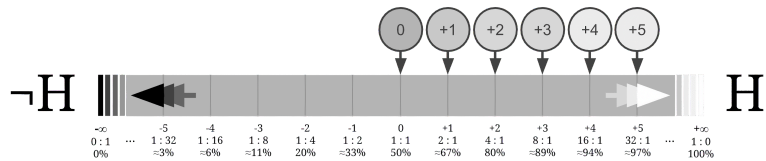
Візуально представити ці зміни ймовірностей можна на осі від 0 до 1:



Ми спостерігаємо, що ймовірності наближаються до 1, але ніколи не досягають того краю: вони просто послідовно долають

частини залишкової відстані, зрештою скупчуючись біля правого кінця.

А якщо перетворити ймовірності у логарифмічні шанси, ситуація стає куди кращою: ймовірності 50% відповідає 0 бітів переконаності, і кожне незалежне спостереження 2: 1 на користь H посуває переконання на одну одиницю вздовж осі:



(Якщо вам цікаво, що відбувається, коли ми досягнемо кінця осі, то такого кінця не існує! 0% імовірності стає $-\infty$ бітів переконаності, а 100% імовірності — $+\infty$ бітів переконаності.)

Інтуїції довкола осі логарифмічних шансів

Є кілька причин, чому ця нескінченна вісь логарифмічних шансів краще узгоджується з нашими інтуїціями про «переконання», ніж звичайний імовірнісний діапазон $[0; 1]$. Наприклад:

- Свідчення, що ви його побачите вдвічі ймовірніше, якщо гіпотеза істинна, — це $^+1$ біт свідчення і $^+1$ біт оновлення, незалежно від того, наскільки впевненим чи невпевненим ви були на початку — адже сила нового свідчення і відстань, на яку ми оновлюємо, не мають *залежати від* нашої апріорної віри.
- Якщо ваша переконаність у чомусь становить 0 бітів — вона ні позитивна, ні негативна, — тоді ваша оцінка шансів становить 1: 1.
- Відстань між 0,01 та 0,000001 набагато більша, ніж відстань між 0,11 та 0,100001.

Зупинімося на останньому пункті: на ймовірнісній осі 0–1 різниця між 0,01 (шанс 1%) та 0,000001 (шанс один на мільйон) згрубша та сама, що й відстань між 11% та 10%. Такий стан справ не узгоджується з нашим інтуїтивним відчуттям сили твердження: різниця між «1 на 100» та «1 на 1 000 000» (!) відчувається набагато істотніше, ніж різниця між «ймовірність 11%» та «ймовірність трохи більш як 10%».

На осі логарифмічних шансів переконаність «1 до 100» відповідає значенню $^{-}2$ порядки величини, а переконаність «1 до мільйона» — $^{-}6$ порядків величини. Між ними відстань у $^{-}4$ порядки величини: $log_{10}(10^{-6}) - log_{10}(10^{-2}) =^{-} 4$ порядки або $^{-}13{,}3$ біта. З іншої сторони, від 11% до 10% — це $log_{10}(\frac{0{,}10}{0{,}90}) - log_{10}(\frac{0{,}11}{0{,}89}) \approx -0{,}954 -^{-} 0{,}907 \approx^{-} 0{,}046$ порядку величини або $^{-}0{,}153$ біти.

Вісь логарифмічних шансів не стискає величезні різниці на краях імовірнісного спектра. Натомість вона демонструє «шкалу переконаності», що нескінченно розширюється в обох напрямках: щоразу, коли ви бачите свідчення зі співвідношенням правдоподібності 2: 1, додається один біт переконаності.

Згідно з законом Вебера — Фехнера[1], більшість людських чуттів логарифмічна за своєю природою: зміна інтенсивності у два рази *відчувається* як збільшення приблизно на ту ж величину, хай би де на шкалі ви стояли. Подвоєння фізичної інтенсивності звуку сприймається людиною приблизно як однакова зміна цього звуку, незалежно від того, чи початкова

1. https://uk.wikipedia.org/wiki/Закон_Вебера_—_Фехнера

інтенсивність звуку становила 40 чи 60 децибелів. Саме тому й існує експоненціальна шкала інтенсивності звуку в децибелах![2]

Тому логарифмічна форма шансів має бути в деякому сенсі *найбільш* інтуїтивним для використання варіантом Баєсової теореми: просто додайте силу свідчень до сили переконання (тобто, якщо зможете зробити свої суб'єктивні оцінки сили свідчення й сили переконання пропорційними логарифмам співвідношень).

Та й за допомогою логарифмічного представлення нам простіше побачити, як надзвичайні твердження потребують надзвичайно сильних свідчень: якщо ваша апріорна віра в H становить -30 бітів і ви бачите свідчення порядку +5 бітів для H, тоді у вас буде -25 бітів віри в H, тобто ви все ще вірите, що гіпотеза менш імовірна за її альтернативи.

Приклад: сині устриці

Розглянемо задачу із синіми устрицями. У Нантáкеті ви збираєте екзотичні устриці. Там є дві різні затоки, де ви шукаєте устриці.

- В обох затоках 11% зібраних устриць містять цінні перлини, тоді як 89% порожні.
- У першій затоці 4% устриць із перлинами — сині, і 8% устриць без перлин теж сині.
- У другій затоці 13% устриць з перлинами сині і 26% без перлин сині.

---

2. На практиці логарифмічне сприйняття звуку означає, що малі зміни добре чутні на фоні тихих звуків (наприклад, шепіт краще розрізнити в порожній кімнаті), а великі зміни майже непомітні на фоні гучних (щоб перекричати рок-концерт і відчути бодай якесь посилення, потрібно колосально збільшити фізичну потужність звуку). Кожні +10 дБ означають зростання фізичної енергії звуку у 10 разів, але суб'єктивно це сприймається приблизно як «удвічі гучніше» — звідси й логарифмічна шкала повзунків гучності в аудіоплеєрах, телефонах та мікшерних пультах: якби вона була лінійною, гучність зростала б надто різко на початку шкали і майже не змінювалася б наприкінці. — Прим. пер.

Ви б, скоріше, взяли синю устрицю з першої чи другої затоки? Для початку зауважимо, що співвідношення правдоподібності від «синьої устриці» до «устриці з перлиною проти устриці без перлини» становить 1: 2 в обох випадках, тому обидва види синіх устриць рівноцінні. (Задумайтеся на мить, наскільки це здавалося б неочевидним до того, як ви дізналися про Баєсову теорему!)

Але який шанс того, що (будь-який з видів) синя устриця ховає в собі перлину? Підказка: саме час перетворити вашу міру переконаності на біти (з основою 2).

> Відповідь 89% — приблизно у 8 разів більше, ніж 11%, тому починаємо з $^{-}3$ бітів віри, що у випадковій устриці є перлина.
>
> Імовірність того, що устриця з перлиною буде синьою, удвічі менша, ніж у порожньої, тому якщо устриця синя, ми отримуємо $^{-}1$ біт свідчення проти наявності перлини. Апостеріорна віра має дорівнювати $^{-}4$ бітам або шансам 1: 16 проти або ймовірності 1/17 (трохи більше за 5% (1/20)). Насправді — 5,88%.

Приклад із життя: тест на ВІЛ

В одному дослідженні донорів крові з Китаю[3] з'ясувалося, що приблизно 1 китайський донор на 100 000 хворів на ВІЛ (значення встановлене дуже надійним тестом, що вважають золотим стандартом). Тести, які використовують для початкового обстеження і не є золотим стандартом, мали *чутливість* 99,7% та *специфічність* 99,8%, що, відповідно, означає $P(\text{позитивний}|\text{ВІЛ}) = 0{,}997$ і $P(\text{негативний}|\neg\text{ВІЛ}) = 0{,}998$, тобто $P(\text{позитивний}|\neg\text{ВІЛ}) = 0{,}002$.

Маємо апріорні шанси 1: 100 000 проти ВІЛ і позитивний результат початкового обстеження на користь ВІЛ зі

3. Приклад суто навчальний і не пов'язаний з дійсними даними. — Прим. пер.

співвідношенням правдоподібності 500: 1.

Застосуймо логарифм з основою 10 (так буде легше розв'язувати в голові):

- Апріорна віра у наявність ВІЛ дорівнює близько -5 порядків.
- Сила свідчення дорівнювала трохи менше за +3 порядки, оскільки 500 менше за 1000 ($log_{10}(500) \approx 2{,}7$).

Отже, апостеріорна віра у наявність ВІЛ трохи нижча за -2 порядки, тобто трохи менше, ніж шанс 1: 100 бути хворим на ВІЛ.

Навіть попри те, що співвідношення правдоподібності для цього обстеження становило 500: 1 на користь ВІЛ, не слід панікувати через позитивний результат!

Певна річ, у цьому дослідженні людей випадково обстежували у країні, де поширеність ВІЛ відносно низька. Для людей, які проходять тестування на ВІЛ через певні побоювання або в країнах із високим рівнем поширеності ВІЛ, потрібні окремі статистичні дані. *Менше з тим*, варто пам'ятати про такі міркування для багатьох хвороб X і спостережень Y, як «лише невелика частина людей хвора на X» та «для попередніх спостережень Y показники хибно позитивних результатів можуть виявитися *не настільки низькими*».

Нескінченні оцінки ймовірності

Представлення у вигляді логарифмічних шансів показує, наскільки значення 0 та 1 є нетиповими у класичних імовірностях. Наприклад, якщо ви присвоїте гіпотезі ймовірність 0 або 1, жодні свідчення не змінять вашої думки.[4]

На осі логарифмічних шансів рівні переконаності коливаються від $-\infty$ до $+\infty$, де нескінченні крайні значення відповідають

4. У статистиці правило Кромвеля (https://defibering.wordpress.com/2023/01/20/scientific-method/#_18._Правило_Кромвеля) стверджує, що емпіричним твердженням не слід присвоювати ймовірність строго 0 або 1, адже завжди існує ймовірність помилитися. — Прим. пер.

0 та 1, які можна розглядати як «нескінченні значення переконаності». Не означає, що 0 та 1 взагалі не допускаються. Раціональний агент виходить з того, що $P(X) + P(\neg X)$ має дорівнювати 1 (де $\neg X$ — логічне заперечення X).[5] У будь-якому разі ці крайні значення нескінченної переконаності поводяться як «особливі об'єкти», а на звичайному спектрі у них якісно інакша поведінка. Такі твердження, як «Після побаченого сильного свідчення моє переконання ніколи не має бути точно таким, яким було», хибні у випадку крайніх значень переконаності, як і «Віднімання 1 від числа дає менше число» хибне, якщо ви наполягаєте на тому, щоб розглядати нескінченність як число.

Свідчення у децибелах

Едвін Джейнз у своїй праці «Теорія ймовірностей: логіка науки»[6] (секція 4.2) пише, що людям легше зрозуміти і використовувати свідчення у децибелах.

Якщо гіпотеза має співвідношення правдоподібності o, тоді його значення у децибелах вираховується формулою $e = 10log_{10}(o)$.

У цій схемі помножити співвідношення правдоподібності на 2 — означає додати приблизно 3 дБ. Помножити на 10 — додати 10 дБ.

Джейнз пише, що через простоту обчислень спочатку використовували десяткові логарифми, а з появою кишенькових калькуляторів намагалися перейти до натуральних логарифмів. Але, як виявилося, легше зрозуміти саме десяткові.

---

5. Нам, простим смертним, краще уникати крайніх імовірностей навіть тоді. — Прим. авт.

6. Jaynes, E. T. *Probability Theory: The Logic of Science (https://archive.org/details/probabilitytheor0000jayn)*. Edited by G. Larry Bretthorst, Cambridge University Press, 2003.

Рухайтеся до найменш неочікуваної гіпотези

Логарифмічна форма шансів Баєсової теореми стверджує: як силу віри, так і силу свідчення можна виміряти в бітах. Цими бітами свідчення так само можна міряти величину під назвою «баєсівське здивування», що дає ще один спосіб інтуїтивно зрозуміти теорему Баєса.

Грубо кажучи, ми можемо оцінити, наскільки гіпотеза H_i дивується свідченням e, визначивши, яку частку ймовірнісної маси гіпотеза надала спостереженню e. Якщо гіпотеза H_i приписала свідченню e 100% своєї ймовірнісної маси, тоді свідчення e є для H_i цілком очікуваним. Якщо ж гіпотеза H_i приписала свідченню e 0% своєї ймовірнісної маси, тоді свідчення e є для H_i *максимально неочікуваним*. Будь-яку міру $P(e|H_i)$ — тобто ймовірність, яку гіпотеза H_i приписує свідченню e, що має цю властивість, можна назвати «здивуванням». Баєсове здивування визначається як $-log(P(e|H_i))$ — величина, що підпорядковується цим інтуїтивним обмеженням і має деякі інші цікаві властивості.

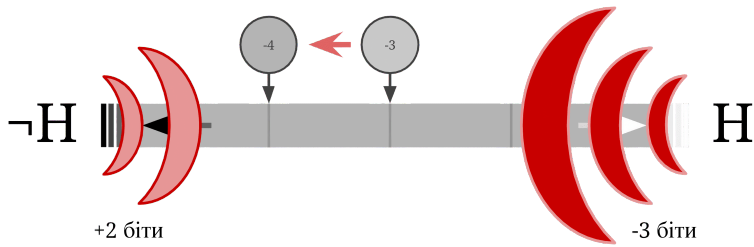
Повернімося до задачі із синіми устрицями. Відповідно, згадаємо і гіпотези H — устриця з перлиною — та $\neg H$ — устриця без перлини. Не будемо ускладнювати собі задачу і припустимо, що ми збираємо устриці вже з третьої затоки, де $\frac{1}{8}$ устриць — ті, що з перлиною, і $\frac{1}{4}$ тих, що без неї, — сині.

Нехай ви знайдете синю устрицю. Гіпотеза H передбачала синій колір у кожної восьмої устриці, тоді як гіпотеза $\neg H$ — у кожної чверті устриць. Отже, для конкретного спостереження e гіпотеза $\neg H$ зробила краще передбачення, оскільки приписала йому вищу ймовірність. Раніше ми сполучали $P(e \mid H)$ та $P(e \mid \neg H)$ в об'єднані співвідношення правдоподібності $\frac{1}{8}:\frac{1}{4} = 1:2$, тобто спостереження ознаки «синій колір» дає 1 біт свідчення для віри в H. А ми ще можемо спершу узяти логарифм, а потім сполучити.

Оскільки H приписала тільки восьму частину своєї ймовірнісної маси спостереженню ознаки «синій колір», а баєсівське оновлення працює за принципом усунення тієї частини маси, яка відповідає результатам, несумісним з отриманим свідченням, ми маємо змістити нашу віру в H на $log_2\left(\frac{1}{8}\right) = -3$ біти від віри в гіпотезу H до $\neg H$. (Кожен негативний біт означає «викинути половину ймовірнісної маси H». А в нашому випадку слід зробити так тричі, щоб виключити те, що H не приписала спостереженню e жодної ймовірності.)

Схожим чином, оскільки гіпотеза $\neg H$ приписала лише чверть своєї ймовірнісної маси спостереженню ознаки «синій колір», ми маємо скоригувати нашу віру в H на $-log_2\left(\frac{1}{4}\right) = -2$ біти від $\neg H$.

Отже, коли отримуємо спостереження ознаки «синій колір», ми посуваємо нашу віру (виміряну в бітах) на 3 поділки від H, потім — на 2 поділки у бік H. У підсумку наша віра посувається на 1 поділку від H.



H приписала $\frac{1}{8}$ своєї ймовірнісної маси до спостереження ознаки «синій колір», тому вона видає $-log_2\left(\frac{1}{8}\right) = 3$ біти здивування, якщо спостереження зсувається від H. $\neg H$ приписала $\frac{1}{4}$ своєї ймовірнісної маси до спостереження, тому вона видає $-log_2\left(\frac{1}{4}\right) = 2$ біти здивування, якщо спостереження відходитиме від $\neg H$ (до H). Загалом, віра в H рухається на 1 біт у бік $\neg H$.

Якби H натомість передбачала ознаку «синій колір» з імовірністю 4% (штраф $log_2(0{,}04) \approx -4{,}64$), а $\neg H$ — з імовірністю 8% (штраф $log_2(0{,}08) \approx -3{,}64$), тоді ми мали б посунути на трохи більше як 4,6 поділки до $\neg H$ і трохи більше як на 3,6 поділки до H, але *сумарно* ми б отримали той самий зсув. Ось чому важлива лише відносна різниця між кількістю бітів, на яку зменшується віра в H, і кількістю бітів, на яку зменшується віра в $\neg H$.

У загальних рисах, за умови спостереження e і гіпотези H, число бітів, які нам потрібно змістити від нашої віри в H, дорівнює $log_2(P(e \mid H))$ — логарифму ймовірності того, що H приписала до e. Ця кількість ніколи не додатна, оскільки логарифм x для $0 \leq x \leq 1$ належить діапазону $[-\infty; 0]$[1]. Якщо змінити їхній знак на протилежний, ми отримаємо невід'ємну величину, що пов'язує H з e, яка дорівнює нулю, коли гіпотеза H надавала цілковиту певність, що e станеться, і нескінченності, коли H надавала цілковиту певність, що e не станеться, а вимірюється в тих самих одиницях, що й свідчення і переконання. Отже, цю величину часто називають «здивуванням», і от, власне, вона вимірює, наскільки гіпотеза H була здивована (у бітах) свідченням e.

Між баєсівським здивуванням і людським відчуттям здивованості (наприклад, коли побачила щось, що, на її думку, було малоймовірним) є деякий взаємозв'язок, але, певна річ, формальне баєсівське здивування і людське емоційне здивування — зовсім різні речі. (Людина може почуватися здивованою і з інших причин, до яких не входить «мої гіпотези не передбачили даних». А ще — люди добре вміють нехтувати свідченнями замість того, щоб відчути здивування.)

З огляду на наведене визначення баєсівського здивування, ми можемо розглядати теорему Баєса як твердження, що *здивування відштовхує віру*. Коли ви помічаєте якесь e, здивування діє

1. Ідеться про діапазон $[0; 1]$, адже $log(0) = -\infty$, $log(1) = 0$. Тому в цьому випадку діє розширена система чисел, в якій існує опція нескінченно великого свідчення від настання майже неймовірної події (з $P = 0$). — Прим. пер.

як сигнал, що відштовхує віру від гіпотези. Повернімося до наведеного вище зображення: коли H передбачає ваше спостереження з імовірнісною масою $\frac{1}{8}$, а $\neg H$ — з імовірнісною масою $\frac{1}{4}$, можна уявити, що H випромінює сигнал здивування силою 3 біти, що відштовхують нашу віру від H, тоді як $\neg H$ випромінює сигнал силою 2 біти, що відштовхують нашу віру від $\neg H$. Обидва сигнали посувають віру в H у різних напрямках, і все закінчується тим, що ви на 1 біт ближче до $\neg H$ (що випромінювала слабший сигнал здивування).

Іншими словами, щоразу, коли ви дивуєтеся якійсь події чи спостереженню, думайте про *пояснення, що має найменше значення баєсівського здивування* для цього свідчення, а потім додайте цій гіпотезі кілька бітів віри.

Теорема Баєса: векторна форма

Шансова форма теореми Баєса стверджує, що для кожної пари гіпотез їхні відносні апріорні шанси, помножені на відносну правдоподібність спостережених даних, дорівнюють відносним апостеріорним шансам.

Нехай H — вектор гіпотез H_1, H_2, ... Оскільки теорема Баєса виконується між кожною парою гіпотез у H, можна просто перемножити вектор шансів на вектор правдоподібності та отримати правильний апостеріорний вектор:

$O(H) \times L_e(H) = O(H \mid e)$,

де $O(H)$ — вектор відносних апріорних шансів між усіма H_i, $L_e(H)$ — вектор відносних правдоподібностей, з якими кожна гіпотеза передбачає спостереження e, а $O(H \mid e)$ — відносні апостеріорні шанси між усіма H_i.

По суті, щоб виконувати кілька оновлень водночас, ми можемо і далі множити на вектори правдоподібності:

$$\begin{aligned} & O(H) \\ & \times L_{e_1}(H) \\ & \times L_{e_2}(H \wedge e_1) \\ & \times L_{e_3}(H \wedge e_1 \wedge e_2) \\ & = O(H \mid e_1 \wedge e_2 \wedge e_3) \end{aligned}$$

Наприклад, у нас є ванна, повна монет. Половина з них — рівноймовірні: всі випадають гербом з імовірністю 50%. Третина монет зміщена в бік герба: гербом випадають у 75% випадків. Решта, навпаки, зміщена на користь іншої сторони: герби випадають у 25% випадків. Ви берете випадкову монету, підкидаєте її тричі й отримуєте такий результат: ЧГЧ. Який шанс того, що ця монета з рівноймовірним випаданням обох сторін?

Маємо три гіпотези: $H_{\text{рівноймовірна}}$, $H_{\text{герб}}$ та $H_{\text{число}}$ з відповідними відносними шансами: $1/2 : 1/3 : 1/6$. Ці три гіпотези приписують відносні правдоподібності 2: 3: 1 того, що монета випаде гербом, для числа співвідношення правдоподібності становить 2: 1: 3. Отже, апостеріорні шанси для всіх трьох гіпотез:

$$\begin{aligned} 1/2 : 1/3 : 1/6 &= 3 : 2 : 1 \\ &\times 2 : 1 : 3 \\ &\times 2 : 3 : 1 \\ &\times 2 : 1 : 3 \\ &= 24 : 6 : 9 = 8 : 2 : 3 = 8/13 : 2/13 : 3/13 \end{aligned}$$

... отже, ймовірність того, що монета з рівноймовірними сторонами, становить $8/13$ або 62%.

Якби ви знали лише ймовірнісну форму теореми Баєса, котру застосовують до однієї гіпотези за раз і лише для обрахунку ймовірності (а відтак і нормалізує шанси в ймовірності на кожному кроці)...

$$P(H_i \mid e) = \frac{P(e \mid H_i)P(H_i)}{\sum_k P(e \mid H_k)P(H_k)}$$

... у вас могли б виникнути зайві складнощі у розв'язанні цієї задачі.

І також, якщо ви чуєте вираз «перевести в шанси, перемножити силу силезну всього, перевести назад у ймовірності» і думаєте: «хммм, схоже на те, що тут перетворення в логарифмічний простір (де усі множники стають доданками) може підвищити ефективність», вітаємо: щойно ви винайшли логарифмічну форму теореми Баєса в шансах. І вона не лише ефективна, а й дає початок природній одиниці вимірювання таких понять, як «сила свідчення» та «сила віри».

Метод наївного Баєса

Множення масиву шансів на масив правдоподібностей — це усталений підхід для створення баєсівських фільтрів спаму. Нехай у нас є три категорії електронної кореспонденції: бізнес, особисте та спам. Останні 100 листів користувач розмітив вручну:

50 — бізнес, 30 — особисте і 20 — спам. Слово «купити» зустрічається у 10 бізнес-листах, 3 з категорії «Особисте» і 10 спамних листах. Слово «раціональність» зустрічається у 30 бізнес-листах, 15 особистих та 1 з категорії «Спам». По-перше, ми припускаємо, що частоти в наших даних репрезентативні для «справжніх» частот. (Буквально: якщо ми бачимо слово, якого ще не бачили, то помножимо його на нульову ймовірність. Краще з цим завданням упоралося б оцінювання частот за методом Ґуда—Тюрінга[1].)

По-друге, ми робимо припущення *наївного Баєса* про зв'язок між наявністю слів «купити» та «раціональність» у тому самому листі: спам-лист зі словом «купити» не більш і не менш імовірно міститиме слово «раціональність», ніж будь-який інший спам-лист.

Тоді ми відфільтруємо повідомлення зі словосполученням «купити раціональність» наступним чином:

Апріорні шанси: 5: 3: 2

Співвідношення правдоподібності для слова «купити»:

$\frac{10}{50}: \frac{3}{30}: \frac{10}{20} = 1\frac{1}{5}: \frac{1}{10}: \frac{1}{2} = 2: 1: 5.$

Співвідношення правдоподібності для слова «раціональність»:

$\frac{30}{50}: \frac{15}{30}: \frac{1}{20} = \frac{3}{5}: \frac{1}{2}: \frac{1}{20} = 12: 10: 1.$

Апостеріорні шанси:

$(5: 3: 2) \times (2: 1: 5) \times (12: 10: 1) = 120: 30: 10 = \frac{12}{16}: \frac{3}{16}: \frac{1}{16}.$

Якщо припущення наївного Баєса відповідають дійсності, апостеріорна ймовірність категорії «Бізнес» становить 75%. А припущення майже напевне *не* відповідають дійсності з причин, які ми опишемо далі. Але хоча розрахунки наївного Баєса зазвичай кількісно хибні, часто вони вказують у правильному якісному напрямку: цей лист імовірніше стосуватиметься бізнесу, ніж якоїсь іншої категорії.

(У практичному застосуванні слід додати логарифмічні правдоподібності замість перемножування співвідношень, щоб

1. https://en.wikipedia.org/wiki/Good%E2%80%93Turing_frequency_estimation

уникнути числового переповнення або замалого значення під час обчислень.)

Ненаївні множинні оновлення

Для виконання множинного оновлення без наївного припущення ми маємо поставити запитання про ймовірність того, що у діловому листі є слово «раціональність» *за умови*, що у ньому було слово «купити».

Наведу приклад з життя: на якомусь із занять з раціональності[2] хтось помітив, як один учасник повів іншого до музею, а другого дня — повів показати своє робоче місце. Невдовзі сформувався ринок ставок довкола того, чи виник між ними романтичний зв'язок. Один учасник стверджував, що, за приблизними підрахунками, ймовірність того, що хтось запросить іншого в музей чи приведе на своє робоче місце, у 12 разів вища, якщо між ними якісь романтичні стосунки, ніж у випадку, коли вони повністю незнайомі. Потім вони перемножили свої апріорні шанси на співвідношення правдоподібності 12: 1 для подорожі до музею та ще одне співвідношення 12: 1 для відвідин роботи і прийшли до висновку, що ті двоє майже напевно були закохані.

Пізніше з'ясувалося, що ті двоє були знайомими з дитинства і в них не було романтичних стосунків. Що ж пішло не так?

Якщо ми хочемо оновлювати гіпотези на основі множинних свідчень, нам слід подумки лишатися в рамках кожної гіпотези та обумовлювати правдоподібність появи майбутніх свідчень за вже спостереженими даними. Припустімо, що ці двоє учасників не перебувають у романтичних стосунках. Бачимо, як вони разом пішли в музей. Якщо подумати, можна дійсно припустити, що ймовірність цієї події становить, скажімо, 1% (зазвичай ми не очікуємо, що незнайомці разом ходять до музею), що може становити близько 1/12 від ймовірності такого спостереження, якби ті двоє були в романтичних стосунках.

2. https://rationality.org/

Але після цього, коли ми спостерігаємо, як він повів її на роботу, ми маємо запитати про ймовірність відвідин місця роботи *за умови*, що ті двоє у романтичних стосунках *і* що вони відвідали музей. Це наводить нас на те, що якщо двоє людей не в романтичних стосунках разом ідуть в музей з якоїсь причини (якою б вона не була), їхня ймовірність спільного відвідування робочого місця не обмежується стандартним показником для пари поза романтичними стосунками. Іншими словами:

$P(\text{робота} \mid \neg\text{стосунки} \wedge \text{музей}) \neq P(\text{робота} \mid \neg\text{стосунки})$.

Метод наївного Баєса, навпаки, наблизить цю кількість $P(\text{музей} \wedge \text{робота} \mid \neg\text{стосунки})$ до множення $P(\text{музей} \mid \neg\text{стосунки}) \times P(\text{робота} \mid \neg\text{стосунки})$. Ось це учасники і зробили, коли двічі помножили на співвідношення правдоподібності $1/12$.

Так вони й припустилися подвійного врахування свідчення: двічі поспіль за два свідчення брали низьку базову ймовірність того, що випадкові люди «кудись ходять разом» і, зрештою, виконали повне оновлення, що виявилося надто сильним.

У фільтрах спаму за методом наївного Баєса часто приписуються абсурдно величезні шанси (порядку гуголів до одного), що якийсь електронний лист — спам чи особистий; і навіть тоді вони інколи помиляються. Якщо в листі є слова «фармацевтичний» та «аптека», фільтр спаму двічі зарахує низьку ймовірність особистого повідомлення зі змістом про ліки, а не врахує те, що якщо я отримаю особистий лист про ліки, він набагато ймовірніше міститиме слово «фармацевтичний». Отже, через припущення в основі методу наївного Баєса наївні баєсівські фільтри спаму аж ніяк не можна назвати такими, чиї ймовірнісні оцінки добре калібровані, до того ж вони надто сильно оновлюють свої апостеріорні шанси на основі отриманих свідчень. З іншого боку, вони часто дають надзвичайно високі оцінки, але в правильному якісному напрямку: лист, якому приписують шанси на те, що він є спамом, у співвідношенні «гугол до одиниці», *не завжди* є спамом, але може виявитися ним, скажімо, у 99,999% випадків.

Виконуючи ненаївні баєсові оновлення на основі множинних свідчень, не забувайте мисленно перебувати у світі, *де гіпотеза істинна*, а потім запитайте, яка правдоподібність кожного наступного свідчення *за умови* істинності гіпотези *та* всіх попередніх спостережень. Не питайте: «Яка правдоподібність того, що пара поза романтичними стосунками піде на чиєсь робоче місце?» — натомість питайте: «Яка правдоподібність того, що пара поза стосунками, яка до того з якоїсь невідомої причини ходила в музей, також піде на чиєсь робоче місце?».

У нашому прикладі з монетами у ванні правдоподібності свідчення на кожному наступному кроці були незалежними — за умови рівноймовірності випадіння обох сторін монети ймовірність випадіння герба під час другого підкидання не залежить від результату першого. Отже, у цьому прикладі припущення в основі методу наївного Баєса було і справді істинним.

Перегляд переконань як вилучення ймовірнісної маси

Механізм дії теореми Баєса можна зрозуміти, зокрема, і через процес оновлення P на підставі нових свідчень, котрий можна тлумачити як вилучення ймовірнісної маси з P (точніше, всю ймовірнісну масу, що суперечить свідченням).

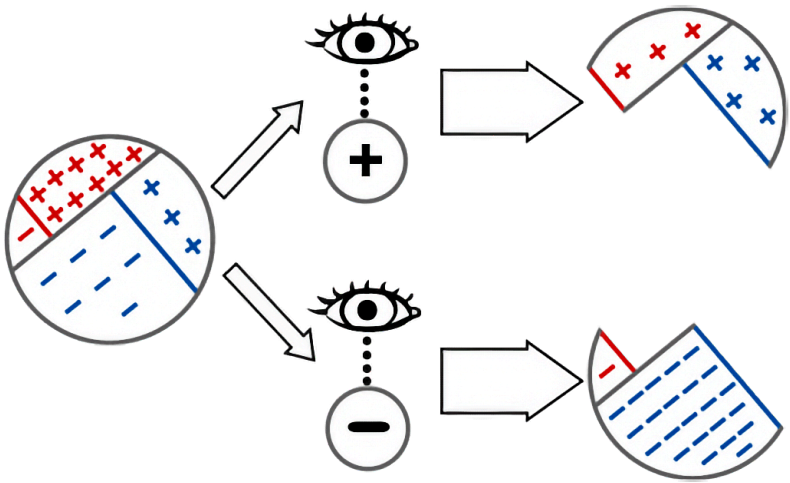
Як приклад використаємо вже відому задачу про дисизіт:

> «Ви обстежуєте якусь множину пацієнтів на наявність хвороби, яку ми назвемо дисизітом. Ви очікуєте, що на дисизіт захворіє близько 20% пацієнтів з обстеженої сукупності. Ви проводите тестування на наявність хвороби за допомогою язикової лопатки із чутливою хімічною смужкою. У 90% пацієнтів, хворих на дисизіт, ця лопатка забарвлюється у чорний колір. Але з тим ця лопатка чорніє і в 30% пацієнтів без дисизіту. До вас у кабінет заходить ваш пацієнт, проходить тестування — його лопатка чорніє. Маючи лише цю інформацію, порахуйте ймовірність того, що у вашого пацієнта дисизіт».

У цій ситуації з одним-єдиним пацієнтом до розгляду яких-небудь свідчень ми можемо перебувати в чотирьох можливих світах:

	Здоровий	Хворий
Тест+	18%	24%
Тест–	2%	56%

Спостерігати, що пацієнт отримує позитивний результат, — означає вилучити з подальшого розгляду можливі світи, де пацієнт отримує негативний результат, і навпаки:



Отже, теорема Баєса стверджує ось що: щоб оновити свої переконання на підставі свідчень, просто відкиньте ймовірнісну масу, що суперечить отриманим свідченням.

Приклад: Задача зі шкарпетками та комодом

Усвідомити, що *спостереження свідчення* відповідає *вилученню ймовірнісної маси*, і зосереджуватися лише на залишковій ймовірнісній масі — ключ до розв'язку задачі пошуку шкарпеток у комоді:

> «Ви кинули свої шкарпетки десь у кімнаті. Вам здається, що шанси того, що ви закинули їх у якусь випадкову шухляду свого комода, становлять ⅘, тому беретеся шукати свої шкарпетки по тих 8 шухлядах комода. Перерivши 6 випадкових шухляд, ви так і не знайшли ні одної шкарпетки. Яка ймовірність того, що ви нариєте шкарпетки у наступній шухляді?».

(За бажанням можете розв'язати задачу, перш ніж читати далі.)

Відповідь: Спочатку у нас є 20% від імовірнісної маси того, що «шкарпетки не в комоді» і 80% від імовірнісної маси того, що «шкарпетки в комоді». Кожним вісьмом шухлядам відповідають 10% ймовірнісної маси (позаяк ймовірність для кожної з восьми шухляд однакова). Вилучивши ймовірнісну масу з 6 шухляд, ми лишаємося з 40% від початкової маси: 20% на те, що «шкарпетки не в комоді», і по 10% на решту непроглянутих шухляд. Оскільки ті залишкові 40% імовірнісної маси — все, що ми маємо, ми розширюємо наш розподіл імовірностей до 100%, тобто ренормалізуємо розподіл імовірностей. Всередині наших залишкових 40% від апріорної ймовірнісної маси половина (початкові 20%) припадає на гіпотезу «не в комоді» і по чверті (початкові 10%) — на ті дві шухляди. Отже, ймовірність знайти свої шкарпетки у наступній шухляді становить 25%. Якщо вам хочеться більш соковитих прикладів застосування методу Баєса, киньте оком на есей «Злети й падіння функції надії в безплідних пошуках[1]».

Розширення суб'єктивної ймовірності

У парадигмі баєсіанства такий підхід *оновлення переконання через обумовлення ймовірнісного розподілу свідченнями* дієвий як у випадках наявності даних про статистичні сукупності з об'єктивними частотностями, що відповідають імовірностям, *так і* в випадках, коли наша невизначеність *суб'єктивна*.

Уявіть себе королем, який думає про непересічну дивачку: ви надаєте близько 20% імовірності, що вона найманий убивця. З цього не випливає, що є сукупність схожих людей, з яких 20% вбивають за гроші. Це означає, що ви зважили свою

1. https://gwern.net/doc/statistics/bayes/hope-function/1994-falk

невизначеність і здогадки та вирішили ставити на те, що ваша підозрювана — найманий убивця, зі співвідношенням 1: 4.

Тоді ви думаєте так: якщо ця дивачка — найманий убивця, з імовірністю 90% вона має при собі кинджал (так ви міряєте власну суб'єктивну невизначеність); якщо ви уявляєте її найманою убивцею, прийнятною ставкою вам здаються шанси 9: 1, що у неї є кинджал. Якщо ця конкретна дивачка не виявиться найманою вбивцею, ваше враження про наявність у неї кинджала становитиме близько 30%.

Коли ваша варта оглядає її і знаходить кинджал, тоді, на думку баєсіанців, ви маєте оновити свої переконання так само, як оновлюєте їх у випадку з дисизітом — де йдеться про велику статистичну сукупність з об'єктивною частотністю захворювання, — попри те, що цей гаданий убивця — унікальний випадок. Якщо вірити баєсіанцям, ваш мозок може відстежувати ймовірності різних можливостей, незалежно від наявності великих статистичних сукупностей та об'єктивних частотностей. Тож коли ви оновлюєте свої переконання за допомогою свідчень, ви не «усуваєте людей з розгляду» — ви вилучаєте *ймовірнісну масу* з конкретних *можливих світів*, представлених у вашому суб'єктивному стані переконань.

Теорема Баєса: форма ймовірностей

Найімовірніше, у підручниках ви стикнетеся із наступним формулюванням теореми Баєса:

$$P(H_i \mid e) = \frac{P(e \mid H_i) \times P(H_i)}{\sum_k P(e \mid H_k) \times P(H_k)}$$

де:

- H_i — гіпотеза, якою ми цікавимося;
- e — отримане нами свідчення;
- $\sum_k$ (вираз, що містить k) — означає «просумуйте вираз за всіма значеннями k»;
- H — множина взаємовиключних гіпотез, що вичерпують всі можливості: H_i — одна з можливих гіпотез, а H_k у цій сумі охоплює всі можливі гіпотези в H.

Нехай у нас є ванна, вщерть заповнена зміщеними монетами.

- Монета першого типу чесна: 50% герб / 50% число. Їх у ванній 40% від усіх монет.
- Монета другого типу видає герб у 70% випадків. Їх у ванній 35% від усіх монет.
- Монета третього типу видає герб у 20% випадків. Їх у ванній 25% від усіх монет.

Нам потрібно дізнатися апостеріорну ймовірність того, що випадково взята монета належить до другого типу після одного підкидання, що дало герб.

Нехай H_1, H_2 та H_3 позначають відповідні гіпотези про монети першого, другого й третього типів. Тепер за допомогою умовних імовірностей нам потрібно обчислити ймовірність $P(H_2 \mid \text{герб})$.

У формі ймовірностей теореми Баєса: $P(H_2 \mid \text{герб}) = \frac{P(\text{герб} \mid H_2) \times P(H_2)}{\sum_k P(\text{герб} \mid H_k) \times P(H_k)}$.

Якщо розкрити суму:

$$P(H_2 \mid \text{герб}) = \frac{P(\text{герб} \mid H_2) \times P(H_2)}{[P(\text{герб} \mid H_1) \cdot P(H_1)] + [P(\text{герб} \mid H_2) \times P(H_2)] + [P(\text{герб} \mid H_3) \times P(H_3)]}$$

.

Підставляємо наші дані в формулу:

$$P(H_2 \mid \text{герб}) = \frac{0{,}7 \times 0{,}35}{[0{,}5 \times 0{,}4] + [0{,}7 \times 0{,}35] + [0{,}2 \times 0.25]} = \frac{0{,}245}{0{,}2 + 0{,}245 + 0{,}05} = 0,(49)$$

.

Рахувати все це було довго й нудно. У цьому немає нічого поганого, адже ймовірнісна форма теореми Баєса добре підходить для безпосередніх обчислень із числами, але не дуже зручна для рахування в голові.

Тлумачення

Ідею теореми Баєса можна сформулювати так: «Подумайте, який кількісний внесок зробила кожна гіпотеза у множині H до нашого очікування побачити свідчення e, включно з правдоподібністю побачити e, якщо гіпотеза H_k істинна, та апріорною ймовірністю H_k. Апостеріорна ймовірність H_i після спостереження e — це внесок H_i у наше очікування побачити e, поділений на сумарний внесок усіх гіпотез».

Якщо говорити точніше, уявіть кожну гіпотезу H_1, H_2, H_3 ... як експерта, який має розподілити ймовірність їхніх передбачень серед усіх можливих свідчень. Конкретніше, «ймовірність» можна уявити як грудку глини. Її загальна маса — один кілограм (імовірність 1). І кожному експерту H_k було призначено частину $P(H_k)$ з того кілограма. Наприклад, якщо $P(H_4) = \frac{1}{5}$, тоді експерту №4 виділили 200 грамів глини.

І разом з експертами ми граємо в гру «Визнач найкращий предиктор».

Щоразу, коли ми збираємося здійснити спостереження E, кожен експерт має розподілити свою глину серед можливих результатів e_1, e_2, ...

Зафіксувавши спостереження $E = e_j$, ми забираємо всю глину, котру поклали не на e_j. Тоді наш новий ступінь довіри до кожного експерта визначається відносною кількістю глини, що в нього залишилася.

Отже, щоб дізнатися, наскільки ми віримо в експерта H_4 після спостереження e_3, нам треба знати дві речі: по-перше, кількість глини, яку H_4 поклав на e_3; по-друге, загальна кількість глини, яку всі експерти (включно з H_4) поклали на e_3.

Своєю чергою, щоб знати *це*, нам потрібно знати, з якою кількістю глини починав експерт H_4 і яку частку своєї глини поклав на e_3. Схожим чином, щоб обчислити сумарну кількість глини на e_3, нам потрібно знати, скільки глини було на початку у кожного експерта H_k і яку частину своєї глини кожен експерт H_k поклав на e_3.

Отже, тут теорема Баєса стверджує наступне:
$P(H_4 \mid e_3) = \frac{P(e_3 \mid H_4) \times P(H_4)}{\sum_k P(e_3 \mid H_k) \times P(H_k)}$.

Які стимули створює ця гра з глиною? На кожному раунді найбільше виграють експерти, які поклали найбільше глини на спостережений e_j, тож якщо ви напевне знаєте, що ось маєте побачити e_3, ви маєте поставити всю свою глину на e_3.

Але ставити *буквально всю* свою глину на e_3 ризиковано — якщо ви спостерігаєте натомість e_5, ви втрачаєте всю глину і вибуваєте з гри. А щойно уся глина експерта зводиться до нуля, він не зможе відновитися у наступних раундах. Гіпотеза більше не може відновитися і, по суті, втрачає вагу. (Дехто називає це «спростуванням».) Якщо ви не певні, що e_5 буквально неможлива, було б мудріше покласти туди хоча б *трохи* глини. Тобто якщо ваш розум наділяє e_5 якоюсь імовірністю, краще вам покласти туди і глину!

(Як виявилося, якщо наприкінці гри ми будемо оцінювати кожного експерта за логарифмом кількості решток глини, тоді на кожному наступному раунді кожен експерт буде вмотивований розміщувати глину точно пропорційно ймовірності, яку він насправді вважає правильною.)

Важливо, щоб експерти заздалегідь розкладали свою глину. Якщо ми дозволимо експертам розкладати глину опісля, вони можуть піддатися спокусі зіграти проти правил, поклавши всю глину на будь-який уже спостережений e_j. Але оскільки ми змушуємо експертів розкласти свою глину заздалегідь, вони мають *розподілити* її на можливі результати: щоб покласти більше глини на e_3, треба забрати її з якогось іншого результату, наприклад, e_5. Щоб поставити дуже високу ймовірність на e_3 й отримати значний відносний ступінь довіри, якщо спостерігається e_3, експерту доведеться ризикнути власною репутацією, якщо замість цього відбудеться якийсь інший результат, наприклад, e_5. Щоправда, це працює лише *за умови*, що ми змушуємо експертів робити прогнози заздалегідь!

З цієї гри ми можемо вивести і те, що відповідь на питання «Чи свідчення e_3 збільшує ймовірність гіпотези H_4?» залежить від того, наскільки добре H_4 передбачила e_3 *порівняно з суперницькими*. Для H_4 недостатньо просто добре передбачити e_3, якщо решта гіпотез так само добре спрогнозувала e_3 — ваша нова чудова теорія фізики отримує нуль очок за передбачення того, що небо синє. Гіпотеза H_k зростає у ймовірності лише тоді, коли вона передбачає e_j краще за альтернативні. І це означає, що ми маємо запитати, що передбачили альтернативні гіпотези, навіть якщо ми думаємо, що ті гіпотези хибні.

Якщо ви потрапляєте в ДТП і не хочете розлучатися з гіпотезою, що ви чудовий водій, тоді ви можете знайти цілу низку причин («дорога була слизькою — от я й не впорався з керуванням!»), чому $P(e \mid$ чудовий водій$)$ не мізерна. Але $P(e \mid$ кепський водій$)$ теж відіграє роль у рівнянні оновлення, позаяк гіпотеза «кепський водій» краще передбачає спостереження ДТП. Тому вашим першим імпульсом, коли ви вирішуєте оновити свої переконання з огляду на ДТП, має бути не «Але моя улюблена гіпотеза дозволяє це свідчення!», а «Нарахувати очки гіпотезі "кепський водій" за краще передбачення порівняно з альтернативами!». (І не забувайте, що вам дозволено трохи

підвищити P(кепський водій) і все ще думати, що вона менше 50%.)

Доведення

Доведення теореми Баєса випливає з визначення умовної ймовірності:

$P(X \mid Y) = \frac{P(X \wedge Y)}{P(Y)}$

та закону повної ймовірності:

$P(Y) = \sum_k P(Y \wedge X_k)$.

Отже:

$P(H_i \mid e) = \frac{P(H_i \wedge e)}{P(e)}$(визначення умовної ймовірності)

$P(H_i \mid e) = \frac{P(e \wedge H_i)}{\sum_k P(e \wedge H_k)}$(закон повної ймовірності)

$P(H_i \mid e) = \frac{P(e \mid H_i) \times P(H_i)}{\sum_k P(e \mid H_k) \times P(H_k)}$(визначення умовної ймовірності)

.

Що й варто було довести.

Баєсівський погляд на наукові чесноти

Через теорему Баєса можна інтуїтивно пояснити низку наукових чеснот, зокрема:

- **спростовність**: доброчесний науковець має сказати, чого він *не* очікує побачити, якщо теорія істинна;
- **сміливість**: влучна теорія продукує сміливі експериментальні прогнози (яких ми б не очікували за інших обставин);
- **точність**: влучна теорія продукує точні експериментальні прогнози (які виявляються правильними);
- **фальсифікаціонізм**: наукова теорія завжди приймається попередньо, а відкидається ж доволі остаточно;
- **експериментування**: збираючи спостереження, ви знаходите кращі теорії, а потім оновлюєте свої переконання.

Спростовність

«Спростовність» означає, що можна вказати, які події та спостереження точно *не* мають відбуватися, якщо теорія істинна.

Уперше цю наукову чесноту популяризував Карл Поппер, який у своїй відомій критиці фройдівського психоаналізу написав таке:

> «Ані Фройд, ані Адлер не виключають того, що конкретна людина може поводитися певним чином, незалежно від зовнішніх обставин. Чи то чоловік пожертвував своїм життям для порятунку дитини від утоплення (випадок сублімації), чи то він убив дитину, втопивши її (випадок витіснення), — обидва випадки не можна ні передбачити, ні виключити фройдівською теорією; *теорія*

просто узгоджувалася з усім, що могло статися»[1].

У рамках баєсіанства спростовність гіпотези ми розглядаємо як вимогу для отримання сильних співвідношень правдоподібності на підтримку гіпотези порівняно з альтернативною, наприклад: «Я не знаю».

Уявіть себе дуже раннім дослідником гравітації на ім'я Ґрек. Ваш друг Таґ тримає в руці камінь, а потім кидає його. Вам треба передбачити, чи той камінь впаде на землю, полетить у небо чи зробить щось інше. Тобто ви маєте сказати, як ваша теорія *Ґрек* приписує ймовірності варіантам «вниз», «вгору» чи «щось інше».

Так сталося, що у вашого друга є своя теорія *Таґ*, згідно з якою «камені роблять те, що хочуть». Якщо Таґ бачить, як камінь падає, він пояснить це спостереження бажанням каменя впасти. Таґ бачить, як камінь рухається вгору, — каже, що камінь хоче полетіти. Таґ думає, що Таґова теорія гравітації дуже вдала, бо нею можна пояснити будь-яку можливу поведінку каменя, яку ми спостерігаємо. Ця особливість робить її кращою за теорію, якою можна пояснити *лише*, скажімо, падіння каменя.

Але Ґрек — баєсіанка і чудово розуміє, що «вниз», «вгору» та «щось інше» — взаємовиключні й вичерпні можливості всередині системи і що *щось* точно має відбутися, коли Таґ дозволяє каменю впасти, — умовні ймовірності $P(\cdot \mid \text{теорія Таґа})$ мусять зводитися до $P(\text{вгору} \mid \text{теорія Таґа}) + P(\text{вниз} \mid \text{теорія Таґа}) + P(\text{інше} \mid \text{теорія Таґа}) = 1$.

Якщо Таґ «однаково добре пояснює» всі три результати — якщо Таґова теорія однаково узгоджується з усіма трьома подіями і видає однаково розумні пояснення для кожного з них, — тоді ми так само можемо сказати, що кожна $P(\text{вгору} \mid \text{теорія Таґа})$, $P(\text{вниз} \mid \text{теорія Таґа})$ та

1. Уривок з Popper, K. (1985). The problem of demarcation (https://todayinsci.com/P/Popper_Karl/PopperKarl-Truth-Quotations.htm). In D. Miller (Ed.), *Popper selections* (pp. 127–128). Princeton University Press.

P(інше | теорія Таґа) дорівнює 1/3. Зауважте, що Таґова теорія (в імовірнісному сенсі) нічим не відрізняється від «Я не знаю».

А тепер уявімо, що Ґрек висуває спростовне передбачення!

Вона каже: «Більшість тіл падає вниз!». Тоді у Ґрек не вся ймовірнісна маса розподілена рівномірно! Ґрек поклала 95% від імовірнісної маси на P(вниз | теорія Ґрек)! І лише 5% рівно розподілила між P(вгору | теорія Ґрек) та P(інше | теорія Ґрек) — у тому випадку, якщо камінь пташкою пурхне в небо. Таґ відказує, що це погана ідея. Якщо раптом камінь полетить вгору, теорія гравітації Ґрек спростується хибним передбаченням! Інше діло — Таґова теорія, що прогнозує шанс 1/3: співвідношення правдоподібності буде 2,5%: 33% ≈ 1: 13 проти теорії Ґрек! Тепер Ґрек спантеличена!

Ґрек каже: вона впевнена, що камінь *впаде вниз*. І птахів куди менше. Тому вона готова ризикнути репутацією і наразитися на можливу помилку. Крім того, набагато важливіше не зберегти лице, а дізнатися, чи теорія Ґрек істинна.

Випускаємо з рук камінь. Камінь падає вниз.

Отримане свідчення має співвідношення правдоподібностей 0,95: 0,33 ≈ 3: 1 на користь теорії Ґрек порівняно з теорією Таґа.

— Як ти отримала таке велике співвідношення правдоподібності? — питає Таґ. — Таґ ніколи не отримував настільки велике співвідношення!

Ґрек пояснює, що вона змогла отримати велике співвідношення правдоподібності, бо піддала ризику свою теорію і *забрала* частину ймовірнісної маси від результатів «вгору» і «щось інше» (якщо щось із цього таки станеться, її теорія буде спростована). Так Ґрек звільняє чималу ймовірнісну масу, яку вона може покласти на результат, що дає велике співвідношення правдоподібності, якщо камінь таки впаде.

Теорія Ґрек виграє, бо вона спростовна і видає коректне передбачення! Якби вона була спростовною і видавала хибне передбачення, то програла б. Але в цьому немає нічого поганого, адже теорія Ґрек — не сама Ґрек.

Спершу передбачення

У наступному експерименті Таґ випускає камінь з рук, дивиться, як той падає, і каже: «Таґова теорія приписує *P*(вниз | теорія Таґа) 100%».

— Ґрек думає, — відповідає вона, — якби Таґ побачив, як камінь натомість злітає вгору, він би сказав, що *P*(вгору | теорія Таґа). Таґ вразливий до упередження заднім числом.

— Ґрек не може довести, що Таґ упереджений, — відказує Таґ, — Ґрек переходить на особистості.

— Нове правило, — каже Ґрек, — усі *заздалегідь* оголошують свої оцінки ймовірностей. А потім вже нема сенсу обговорювати.

Таґ думає:

— Таґ каже *P*(вгору | теорія Таґа) = 1 і *P*(вниз | теорія Таґа) = 1.

— Таґ порушує аксіоми теорії ймовірностей, — каже Ґрек, — ймовірність усіх взаємовиключних результатів мусить дорівнювати одиниці або меншому значенню. Але добре, що Таґ сказав це зараз і Ґрек бачить проблему.

— Так нечесно! — заперечує Таґ. — Має бути дозволено оголошувати оцінки після події — тоді ніхто не зможе довести, що я помилився!

Формальність

Правило «спершу передбачення» набагато важливіше у прагматичному плані для неформальних теорій, аніж для формальних; у цьому контексті теорія вважається «формальною», якщо її передбачення формуються настільки механічно й однозначно, що будь-хто може ввести її в комп'ютер і отримати той самий результат у ймовірностях, який припускає ця теорія.

Коли Ньютон запропонував свою теорію гравітації, її вважали ще не до кінця доведеною, позаяк такі ретропрогнози (передбачення вже відомих результатів), як припливи, еліптичні орбіти планет і закони Кéплера[23], вже можна було спостерігати ще *до* Ньютона.

Навіть у такому випадку баєсіанець надав би Ньютоновій теорії значну вагу завдяки цим ретропрогнозам, бо, на відміну, скажімо, від психологічної теорії людської поведінки, будь-хто — а не тільки Ньютон — міг узяти олівець і вивести точно такі ж передбачення з Ньютонових законів. Не можна повністю виключати ту можливість, що Ньютонову теорію в якомусь сенсі надміру підлаштували під закони Кеплера і дані про припливи, внаслідок чого вона не здатна буде давати нові коректні передбачення. Але менше практичних підстав побоюватися, що Тагтон просто скаже: «Ну, звичайно ж, моя теорія "Планети рухаються, куди хочуть" передбачала б еліптичні орбіти; еліптичні орбіти виглядають класно».

Запитати послідовників теорії, що вона стверджує

Таґ піднімає ще один камінь:

— Я наперед кажу, що теорія Ґрек приписує ймовірність 0%, що камінь впаде, — Таґ спростовує теорію Ґрек!

Ґрек хитає головою:

— Треба запитати пропонентів теорії Ґрек, що вона передбачає, — Ґрек піднімає інший камінь. — Я кажу, що теорія Ґрек приписує P(вниз | теорія Ґрек) = 0,95.

— А я кажу, теорія Ґрек стверджує P(вниз | теорія Ґрек) = 0 , — заперечує Таґ.

---

2. https://uk.wikipedia.org/wiki/Закони_Кеплера

3. Три закони, сформульовані Йоганном Кеплером, що описують форму планетних орбіт, зміну швидкості руху планет та зв'язок між розміром орбіти й періодом обертання. В цьому прикладі закони Кеплера є раніше відомими спостережуваними закономірностями, які Ньютонова теорія могла пояснити та вивести з більш загальних принципів. — Прим. пер.

— Наука так не працює, — відповідає Ґрек, — Таґ має сказати, що стверджує Таґова теорія.

На мить Таґ задумується:

— За теорією Таґа, камінь впаде вниз з імовірністю 95%.

— Що? — дивується Ґрек. — Таґ просто бере теорію Ґрек! І Таґ казав не це *до того*, як побачив падіння каменя!

Таґ лукаво хмилиться:

— Тільки ж Таґ має сказати, що передбачає його теорія. Правда ж?

Знову ж — з практичних цілей *спершу* нам слід запитати послідовників неформальної теорії, що ця теорія передбачає (формальну теорію може застосувати будь-хто; якщо це не так, «формальною» її назвати не випадає).

Більше того, з тієї причини, що будь-який дурень може слідувати будь-якій справі, вам слід запитати найрозумніших та найбільш технічно підкованих пропонентів цієї теорії. Якщо виникає спір щодо того, кого таким вважати, запитайте окремо представників цієї спільноти, котрі вважаються найкомпетентнішими. Слава — точно не основний маркер: питання з квантової механіки слід ставити Мюррею Ґелл-Манну[4], а не Діпаку Чопрі[5], навіть якщо більше людей чули думки другого, а не першого. Якщо ви справді не можете провести різницю між ними, запитайте їх обох — не питайте лише Чопру, щоб потім сказати, що Чопра і є *тим* представником, бо про нього всі кажуть.

У такого роду реверансах не було б потреби, якби у нас був досить розвинений штучний інтелект чи досконалий раціональний агент, проте у людському наукотворенні, де люди можуть бути вмотивовані хибно трактувати ймовірнісні оцінки іншої теорії, такі правила мають сенс.

У цього неформального правила є обмеження, і можуть бути випадки, коли здається очевидним, що передбачення гіпотези не збігаються із заявами її послідовників або вони просто крадуть передбачення більш вдалої теорії. Але мусить існувати *значна*

4. https://uk.wikipedia.org/wiki/Маррі_Гелл-Манн

5. https://nashformat.ua/authors/dipak-chopra-books

(хоча може бути скасована за наявності підстав) презумпція на користь того, щоб прихильники теорії могли висловлювати те, що, згідно з цією теорією, має відбутися.

Сміливість

Через кілька хвилин Ґрек піднімає ще один камінь:

— P(вниз | теорія Ґрек) = 0,95.

— P(вниз | теорія Таґа) = 0,95, — каже він, — бачиш: Таґ приписує вищу ймовірність спостереженим результатам. Таґ виграє?

— Ні, — відповідає Ґрек, — співвідношення правдоподібності 1:1 тепер весь час, навіть якщо ми віримо Таґу. Таґова теорія не отримує переваги. Таґу треба висловити *сміливе* передбачення, якого не роблять інші теорії.

Таґ насуплює брови:

— Таґ каже... камінь стане синім, коли ти випустиш його з рук? P(синій | теорія Таґа) = 0,9.

— Дуже сміливо, — сказала Ґрек, — теорія Ґрек не стверджує такого (ані якийсь очевидний «здоровий глузд», ані теорії «як зазвичай»). Ґрек вважає, що P(синій | ¬ теорія Таґа) < 0,01, тому Таґове передбачення точно можна назвати сміливим. Якщо воно справдиться, ця подія буде справді вагомою.

$$\frac{P(\text{теорія Таґа} \mid \text{синій})}{P(\neg\text{теорія Таґа} \mid \text{синій})} > 90 \times \frac{P(\text{теорія Таґа})}{P(\neg\ \text{теорія Таґа})}$$

— Тепер Таґ виграв? — питає він.

Ґрек випускає камінь. Падає на землю. Не стає синім.

— Сміливе передбачення хибне, — відповідає Ґрек, — Таґове передбачення доброчесне, але не перемагає. Тепер Таґ програє співвідношення правдоподібності 1:10. Ось тобі наука, ось тобі спростовність.

— Ґрек заманює Таґа в пастку! — обурюється Таґ.

— Дивись, — каже Ґрек, — вся заковика у тому, щоб встановити правила, за яких перемагають *правильні* теорії. Якщо конструктори хибних теорій, намагаючись поводитися як доброчесні науковці, швидко програють, це перевага — не вада. Але

якщо Таґ пробує бути доброчесним і програє, ми тиснемо руки, і Таґ все ще виглядає добре. Зрештою, таким і має бути нормативний соціальний ідеал.

Точність

На якомусь пізнішому щаблі розвитку теорії гравітації арістотелівський синтез теорій Ґрек і Таґа — «Більшість тіл мають кінцевий пункт призначення — центр Землі — і намагаються досягти його» — вступає у двобій з теорією Ґалілео Ґалілея — «Більшість тіл у вільному падінні прискорюються вниз, і з кожною секундою швидкість тіла збільшується на 9,8 м/с; не питайте чому — кажу, що бачу».

— Ти ж лише передбачаєш те, що і так видно: якщо камінь упустити, він падатиме вниз, — каже Арістотель, — я ж уже це передбачив.

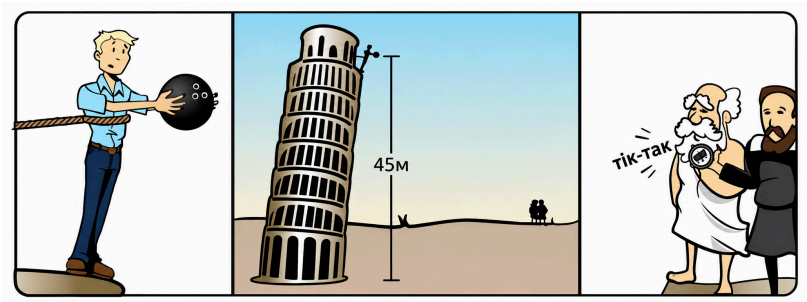
— Наступним кроком, — починає Ґалілео, — ми передбачаємо, *як довго* куля для боулінгу падає з Пізанської вежі. — Ґалілео дістає з кишені секундомір: — Щойно мій друг скине кулю, ти тиснеш кнопку пуску, а як куля досягне землі, тиснеш кнопку зупинки. І ми подивимося, які саме числа покаже нам секундомір.

Після додаткової перевірки, чи в Арістотеля стабільний час реакції на натискання кнопки на секундомірі, коли Ґалілей клацає пальцями, Арістотель піднімає погляд на кулю, яку тримають на вершечку Пізанської вежі.

— Думаю, куля для боулінгу буде падати від 0 до 5 секунд включно, — прогнозує Арістотель, — а далі вже не впевнений.

— Гаразд, — відповідає Ґалілео, — я поміряв висоту вежі — 45 метрів. Отже, якщо опір повітря дорівнює нулю, після 3 секунд куля має падати зі швидкістю $9{,}8 \times 3 = 29{,}4$ м/с. Ця швидкість безперервно зростає протягом цих трьох секунд, тож *середня* швидкість кулі становить $29{,}4/2 = 14{,}7$ м/с. А якщо за три секунди куля рухається із середньою швидкістю 14,7 м/с, вона подолає шлях униз на 44,1 метра. Отже, куля падатиме трохи

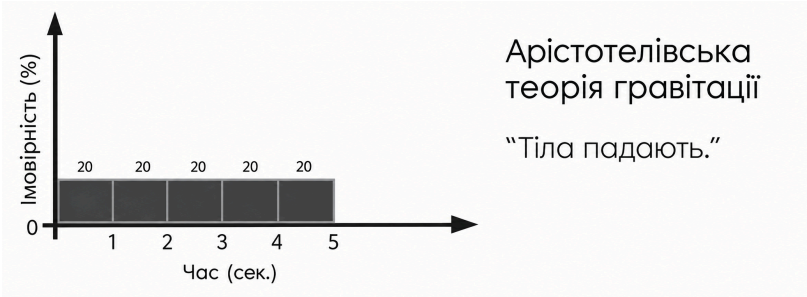
довше, ніж три секунди, щоб подолати шлях у 45 метрів. Щось близько до додаткових 1/29 секунди.



— Хммм, — задумується Арістотель. — Твій кишеньковий годинник вимірює час у секундах. Тож твоя теорія вкладає усю ймовірнісну масу у три секунди. Правильно?.

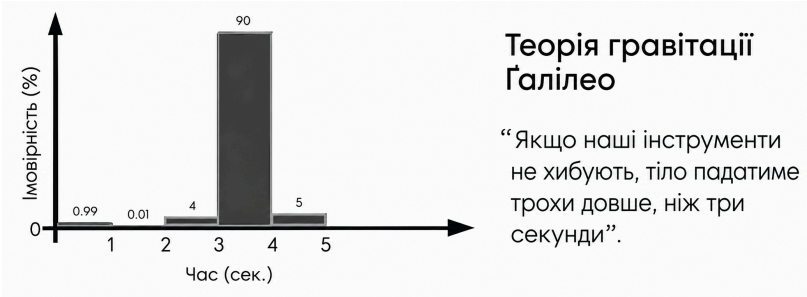
— Не прямо *всю* масу, — відказує Ґалілео, — ще проходить якийсь час перед тим, як ти побачиш падіння кулі й натиснеш кнопку пуску, а ще ж якийсь час мине між вдаряннням кулі об землю і натисканням кнопки зупинки. *Здебільшого* ці два джерела похибок у вимірюваннях мають компенсуватися, але, може, саме в цьому випадку так не буде. У нас тут немає достатньо точної й добре перевіреної постановки експерименту. Тобто, якщо секундомір зламається і ми побачимо «0» на циферблаті, теорія гравітації Ґалілео зазнає *поразки*, але *остаточного* спростування чекати не варто — можна роздобути іншого годинника, сформулювати краще передбачення і надолужити цю поразку наступними результатами.

— Гаразд. Які ймовірності тоді приписує твоя теорія? — питає Арістотель. — Гадаю, моя теорія приблизно однаково добре пояснює будь-який час падіння від 0 до 5 секунд.



Ґалілео роздумує:

— Оскільки ми ще не тестували таку постановку... Думаю, поставлю десь 90% моєї ймовірнісної маси на те, що час падіння буде між 3 і 4 секундами, що відповідає спостережуваному результату на годиннику, що показує «3». Може, покладу ще 5% імовірності на опір повітря, що може мати відчутніший ефект, ніж мені здається, на цій дистанції, тому між 4 і 5 секундами, якщо циферблат покаже «4». І ще 4% на те, що годинник повільніший, ніж я думав, тому 4% ставлю на те, що час буде між 2 і 3, а секундомір покаже «2». І дам 0,99% ймовірності того, що годинник зламається і покаже «1» (або «0», але ми обоє погодилися, що такого не станеться), і 0,01% на те, що я побачу «2», чого взагалі-то не має статися, як я думаю.



— Що ж, — каже Арістотель, — безперечно, у твоєї теорії є наукова чеснота *точності*: зосереджуючи більшу частину ймовірності у вузькому діапазоні можливих точних спостережень, ви заробляєте своїй теорії вищу правдоподібність, що дає перевагу над *розпливчастішими* теоріями, як моя, яка стверджує лише «падіння тіл» і дає «справджене передбачення» щоразу, коли тіла падають вниз, але не передбачає, як довго вони будуть падати. Якщо твоє передбачення «3» справдиться, співвідношення правдоподібності становитиме 0,9: 0,2 або 4,5: 1 на користь теорії Ґалілео проти теорії Арістотеля.

— Саме так, — підтверджує Ґалілео, — певна річ, передбачення має бути не лише точним, а й правильним. Якщо годинник натомість показує «4», співвідношення правдоподібності становитиме 0,05: 0,2 або 1: 4 проти моєї теорії. Краще висунути розпливчасту й правильну теорію, ніж точну й помилкову.

Арістотель киває:

— Тоді до експерименту.

Боулінгову кулю кинуто.

Секундомір показує 3 секунди.

— То тепер ти віриш моїй теорії? — питає Ґалілео.

— Ну, я вірю їй десь у 4,5 раза більше, ніж до цього, — відказує Арістотель. — Але та частина, де ти підставляєш числа штибу «9,8» і всі ті обрахунки типу піднесення часу до *квадрата*, здається мені складнуватою. Тобто, якщо мені можна підставляти

такі вже точні числа і виконувати такі операції, як піднесення до квадрата, мусять бути сотні різних теорій, які можуть бути такими ж складними. За допомогою чисельної форми леза Оккама[6] нам потрібно ще штрафувати апріорну ймовірність твоєї теорії за її алгоритмічну складність[7]. Одного спостереження зі співвідношенням правдоподібностей 4,5: 1 недостатньо, щоб виправдати всю ту складність. Я не збираюся вірити у щось таке складне лише тому, що раз побачив трійку на секундомірі! Я хочу побачити, як з різних висот падають різні тіла, і пересвідчитися, що отримані значення часу відповідають твоїм передбаченням. Якщо я скажу, що апріорна складність твоєї теорії — нехай 20 бітів, тоді ще додаткових 9 подібних спостережень переконають мене. Звісно, я очікую, що ти вже зробив більше спостережень на самоті, але загальнодоступним знанням людства воно стає тоді, коли хтось відтворює його.

— Звісно, — відповідає Ґалілео, — мені б хотілося перевірити твою постановку експерименту, а особливо твої розрахунки, коли ти спробуєш їх провести кілька перших разів, щоб упевнитися, що ти міряєш у метрах, а не в футах, чи не забув поділити кінцеву швидкість навпіл для отримання середньої швидкості і таке інше. Хоч ця теорія і формальна, хочу перевірити, щоб ти не зробив помилки в розрахунках.

— Ясна річ, — підтверджує Арістотель. — Який же я все-таки радий, що ми обоє баєсіанці і знаємо основні закони теорії

6. Ідеться про індукцію Соломонова — формалізований підхід до індуктивного висновування, у якому простішим гіпотезам, тобто гіпотезам із коротшим алгоритмічним описом, надається більша апріорна вага. У контексті цього есею це формалізація леза Оккама: коли дві гіпотези однаково добре пояснюють дані, простіша з них має вищу апріорну ймовірність. — Прим. пер.

7. Алгоритмічна, або колмогоровська (https://prepost.wikidot.com/kolmogorovska-skladnist), складність — це довжина найкоротшої програми, яка може відтворити певну інформацію на заданій універсальній машині. Що коротше можна описати теорію або модель, то меншою є її алгоритмічна складність. Тут ідеться саме про таке формальне розуміння «простоти»: складність теорії можна оцінювати за довжиною найкоротшого алгоритмічного опису, потрібного для її відтворення (див: колмогорови (https://prepost.wikidot.com/pirsony-i-kolmogorovy)). — Прим. пер.

ймовірностей і те, як вони скеровують неформальні соціальні процедури, яким ми слідуємо.

— Як з язика зняв, — і собі говорить Ґалілео. — Інакше ми б уже загрузли у затятому срачі, що міг би затягнутися на години.

Фальсифікаціонізм

Однією з причин зачарованості Карла Поппера «спростуванням» було спостереження, що в науці спростування визначеніше й остаточніше за підтвердження. Класична притча в цьому ж дусі — Ньютонова теорія гравітації проти загальної теорії відносності Айнштайна: попри гори експериментальних свідчень на користь теорії Ньютона, визбируваних аж до XIX століття, не було жодного сенсу, в якому можна було б сказати, що Ньютонову теорію гравітації *остаточно підтверджено*, та з тим її таки відкинули на користь Айнштайнової. І вже тепер, коли Ньютонову теорію викинули на смітник історії науки, немає ніякої реалістичної ймовірності, що *колись* її повернуть назад — на відміну від прийняття, відкидання остаточне.

Оскільки Поппер працював до тих часів, коли теорія Баєса набула широкої слави, він надавав цій схемі *логічної* інтерпретації. Нехай $H \rightarrow E$, де гіпотеза H логічно передумовлює спостереження свідчення E. Якщо замість цього ми спостерігаємо $\neg E$, за принципом контрапозиції[8] ми можемо вивести $\neg H$. З іншого боку, якщо ми спостерігаємо E, то вже не випадає логічно виводити H. Тому ми можемо логічно спростувати теорію, але не логічно підтвердити.

На ділі наука часто так не працює. У XIX столітті під час спостережень за орбітою Урана накопичилися аномалії. І навіть після врахування усіх відомих впливів від інших планет Уран рухався все ще не *точно* так, як диктувала Ньютонова теорія. З погляду логіко-фальсифікаційного, оскільки, за Ньютоновою теорією гравітації, Уран має бути у певній точці, де його не було, ми повинні стати нескінченно певними, що Ньютонова теорія

8. https://uk.wikipedia.org/wiki/Контрапозиція

хибна. Дехто з теоретиків припустив, що у Ньютонової теорії може бути якийсь маленький член похибки, з якого випливає і хибність її початкової форми.

У дійсності Урбен Левер'є[9] та Джон Кауч Адамс[10] незалежно один від одного припустили, що аномалія на орбіті Урана може бути спричинена раніше не спостережуваною восьмою планетою. І замість того, щоб *нечітко* заявити, що такою була їхня гіпотеза (такий крок, по суті, розподілив би ймовірнісну масу положення Урана ширше й тим самим зробив би спростування ньютонівської механіки *менш вираженим*), мосьє Левер'є та містер Адамс незалежно один від одного взялися обчислювати очікуване положення восьмої планети. У 1846 році, виходячи з розрахунків мосьє Левер'є, Йоганн Ґалле спостеріг Нептун. У той день ньютонівська механіка тріумфувала.

У 1859 році Урбен Левер'є помітив ще одну проблему: цього разу очікуване положення Меркурія не збігалося зі спостережуваним. Тоді як Ньютонова теорія гравітації передбачала, що орбіта Меркурія має піддаватися прецесії — точка найближчого зближення з Сонцем має повільно обертатися навколо Сонця, — прецесія Меркурія перевищувала норму на 38 кутових секунд щостоліття (пізніше це значення скоригували до 43). Цю аномалію було вже складніше пояснити; мосьє Левер'є вважав, що навколо Сонця — всередині орбіти Меркурія — обертався якийсь крихітний планетоїд[11].

Трохи більше ніж через півстоліття герр Айнштайн, працюючи над рівняннями Загальної теорії відносності, усвідомив, що аномальну прецесію Меркурія цілком пояснювали його рівняння у своїй найпростішій та найелегантнішій формі.

На цьому Ньютоновій теорії гравітації прийшов кінець. Назавжди.

Якщо подивитися на ситуацію з логічного погляду Поппера, між аномалією Урана та аномалією Меркурія немає ніякої

9. https://uk.wikipedia.org/wiki/Урбен_Левер%27є

10. https://uk.wikipedia.org/wiki/Джон_Кауч_Адамс

11. https://uk.wikipedia.org/wiki/Вулкан_(планета)

очевидної різниці. В обох випадках прямолінійне ньютонівське передбачення здавалося спростованим. Якщо Ньютонова теорія гравітації змогла відновитися від одного логічного спростування, то чому б їй не оговтатися після спростування? З погляду баєсіанства, можна побачити різницю наступним чином.

У випадку Урана кращої альтернативи ньютонівській механіці, що давала б правильніші передбачення, не було. Здавалося, тодішня теорія явно помилялася щодо Урана: ньютонівська модель робила впевнені передбачення щодо положення Урана, які виявлялися значно помилковішими, ніж теорія мала б очікувати в середньому. Це веде нас до того, що має існувати *якась* краща альтернатива і не обов'язково, щоб вона була неньютонівською. Низька P(розташування Урана | нинішня ньютонівська модель) створила потенційну можливість отримати сильніший прогноз за модифікованої моделі із вищою P(розташування Урана | нова модель), хоч і не казала, що саме має змінитися у новій моделі.

І навіть після відкриття Нептуна ньютонівська механіка не була *остаточно* підтверджена. Хоча нова модель передбачала високу P(розташування Урана | Нептун ∧ Ньютонова теорія), наскільки можна було знати, могла існувати якась невідома Інша теорія, що передбачала б точно таку ж високу P(розташування Урана | Нептун ∧ Інша теорія). У такому випадку Ньютонова теорія не мала б переваги за правдоподібністю над цією невідомою Іншою теорією, тож ми не можемо сказати, що Ньютонова теорія гравітації була підтверджена серед *кожної іншої можливої* теорії.

У випадку Меркурія, коли з'явилася Айнштайнова формальна теорія і передбачала куди вищу P(розташування Меркурія | Айнштайнова теорія), ніж P(розташування Меркурія | Ньютонова теорія), вона заробила собі величезне співвідношення правдоподібності і занизила ймовірність Ньютонової теорії. Навіть якщо одного дня якась інша теорія виявиться кращою за Айнштайнову — тобто даватиме ту

саму *P*(розташування Меркурія | Інша теорія) і навіть кращу *P*(нове спостереження | Інша теорія), — те, що Айнштайнова теорія показала набагато кращі результати по Меркурію, свідчить про те, що прості теорії *можуть* давати нам кращі передбачення руху Меркурія простим способом, що точно не є ньютонівським. Отже, якою б не була та нова Інша теорія, вона також даватиме кращі передбачення щодо Меркурія, ніж *P*(розташування Меркурія | Ньютонова теорія), у неньютонівський спосіб; зрештою, Ньютонова теорія все одно програє тій Іншій теорії з величезною перевагою за правдоподібністю щодо Меркурія.

Отже, з погляду баєсіанства, після пояснення орбітальної прецесії Меркурія ми не можемо бути певні того, що Айнштайнова теорія гравітації правильна, але *можемо* бути певні, що Ньютонова теорія гравітації хибна.

Але і цей опис не відображає логічної різниці між спростуванням та підтвердженням: усе відбувається у світі ймовірностей.

Можливість постійного підтвердження

Варто зазначити: хоча Ньютонова теорія гравітації була хибною, щось *дуже подібне до неї* було істинним. Тобто, хоча твердження «Планети рухаються *саме* так, як казав Ньютон» можна було прийняти як тимчасове і врешті-решт його спростували, твердження «Усі типи планет, які ми досі бачили, в тих умовах, яких ми за ними досі спостерігали, рухаються *практично* так, як передбачає Ньютонова теорія гравітації» підтверджувалося набагато краще.

З цього випливає, що, попри відкидання Поппером самого поняття підтвердження, *деякі* теорії можна підтвердити поза межами обґрунтованого сумніву. Наприклад, теорія ролі ДНК у біологічному відтворенні. Що б ми не думали про кварки, немає ніяких обґрунтованих підстав вважати, що ми можемо помилятися щодо існування молекул чи наявності молекул у формі

подвійної спіралі, що кодує генетичну інформацію. Є підстави вважати, що теорію ДНК підтверджено назавжди і поза межами розумних сумнівів, тому вона ніколи не піде на смітник історії науки, хай би які нові спостереження з'явилися.

Таке можливо, бо ДНК — нефундаментальна теорія, якщо говорити про неї у термінах молекул та атомів, а не кварків. Навіть якщо кварки не точно те, що ми думаємо, буде існувати щось *схоже* на кварки, що лежатиме в основі об'єктів, які ми звемо протонами та нейтронами, існуватимуть теж атоми і молекули вже на вищому рівні, а це означає, що об'єкти, які ми називаємо ДНК, усе ще існуватимуть у новій теорії. Іншими словами, біологічна теорія ДНК — це теорія, що *в своїй основі* містить «щось *як оце*, що мусить бути істинним». Ця гіпотеза, за якою те, що Джозеф Блек називав «фіксованим повітрям»[12], а ми називаємо «діоксидом вуглецю», насправді зроблене з одного атома карбону і двох атомів оксигену, була назавжди підтверджена, як Ньютонова теорія гравітації назавжди спростована.

Існує певна кількість спостережень, що могли б переконати нас, що вся наука — брехня, а в саду живуть феї, але, менше з тим, вуглекислий газ нікуди не зникне.

Проте у звичайній науці, коли ми намагаємося розібратися у суперечливих питаннях, застосування теореми Баєса веде до того, що доброчесний науковець має мислити так, як пропонував Карл Поппер:

- вважати спростування сильнішим за підтвердження;
- приймати гіпотези з великою кількістю на перший погляд сприятливих свідчень лише тимчасово;

12. Історична назва вуглекислого газу. У середині XVIII століття Джозеф Блек досліджував луги та карбонати й установив, що в деяких твердих речовинах міститься особлива газоподібна речовина, яку він назвав «фіксованим повітрям». Вона могла сполучатися з твердими речовинами, наприклад із вапном, а під час нагрівання або дії кислот вивільнялася з них. Пізніше цю речовину досліджував Антуан Лавуазьє й установив, що вона є сполукою вуглецю з киснем. — Прим. пер.

- мати якусь кількість заперечувальних свідчень і невдалих передбачень, внаслідок яких гіпотеза потрапляє на смітник історії науки і відмовляється від надії на її воскресіння;
- перш ніж вважати теорію «підтвердженою», вимагати *якісно* сильнішого типу свідчень — безпосереднього спостереження складників і процесів явища в подробицях.

Експериментування

Якою б не була правдоподібність *P*(спостереження | гіпотеза), вона не змінює ваших переконань, якщо ви не проводите експеримент і не дізнаєтеся, чи відбулося передбачене спостереження, чи ні, і не оновлюєте свої переконання з урахуванням отриманого результату.

У цьому сенсі також можна сказати, що теорема Баєса спонукає до експериментального методу, хоч вам і не обов'язково знати багато математики, щоб усвідомити, що для накреслення точної мапи міста вам потрібно дивитися на це місто. Та все ж, позаяк протягом великої частини історії людства експеримент не був аж таким очевидним методом, йому б не завадила будь-яка підтримка, зокрема, центральна ідея баєсіанства: «Проводьте спостереження, щоб оновлювати переконання».