English Translation
Hi everyone! I wanted to share my thoughts.

IN BRIEF: THE CORE IDEA
Today's neural networks are genius compilers. They have read the entire library of humanity and can write a brilliant essay on any topic, but they have never left this library. They cannot conduct their own experiments or make discoveries that aren't already in the books. They are locked within the confines of what we, humans, already know.
The proposed idea:

1. **Intrinsic Motivation:** The driving force of this process should be "structural tension"—an analog to discomfort or curiosity that arises when encountering a contradiction. Instead of just responding to prompts, it would experience an internal "tension" from contradictions in its worldview. This drive for coherence and order would become its engine.
2. **Top-Down Approach:** Before speaking, it must first form a holistic "intent" or "blueprint" for the answer. Only then would it select the words to express that intent. This is like an architect who first creates the building's design and only then chooses the bricks.
3. **Growth Cycle:** The AI must learn to formulate its own hypotheses (abstractions), test its ideas in the real world, discard incorrect ones, and turn validated hypotheses into its foundational "beliefs."

Ultimately, we would get not an echo of human knowledge, but an independent mind capable of genuine discoveries.

A MORE DETAILED EXPLANATION
They don't understand *why* data is important, they don't perceive time, and they cannot independently set goals beyond those given by humans. If the data is contradictory, the probability of an erroneous choice increases sharply. They have no value system of their own, and so on. Imagine two types of intelligence.
Type 1: The "Erudite in a Locked Room" (Today's LLMs) This intelligence has read absolutely all books, articles, and websites. It sees that the word "sky" often appears next to "blue," and "fire" next to "hot." In response to any question, it can assemble a statistically probable and smooth answer from fragments of what it has read.
The Problem: It doesn't understand *why* the sky is blue. It has no concept of light, atmosphere, or scattering. If it encounters a contradiction or a new problem, it has no way to find out the truth. It can only compile the most probable, but potentially false, answer (these are its "hallucinations"). It operates on a "bottom-up" principle: word by word, brick by brick, without an initial blueprint for the entire building.
Type 2: The "Child-Experimenter" (The Proposed AGI) But where would the internal motivation for self-improvement come from? What would be the driving force of this process? The breakthrough to true Artificial General Intelligence (AGI) is not about accumulating more data, but about igniting an inner fire. This intelligence doesn't just read books. Its main goal is to build a consistent and coherent model of the world within itself.
1. The Inner Engine: An "Instinct for Truth." When this AI encounters a contradiction—either with its previously learned data or with an observation (e.g., its model says a stick is straight, but in water, it sees it as bent)—it doesn't just trigger a logical error, but a kind of "pain" or "structural tension." This is an internal, innate motivation to resolve the conflict and restore the integrity of its worldview. This is its main driving force, compelling it to learn, becoming the fuel for curiosity and the primary engine of its development.
2. The Birth of Intent: "Essence First." This internal tension creates a key inversion in the thought process. Current LLMs build answers from the bottom up: from word to word, feeling out the most probable path. The proposed AGI would work from the whole to its detailed expression. The internal

conflict gives rise to a "need." The need forms an "intention" to resolve it. The intention creates a holistic, yet undetailed, "conception" (a blueprint)—the hypothesis itself. To resolve this contradiction, the AI doesn't start generating random words. It works "top-down": * First, it forms an abstract hypothesis, an "intent" or "blueprint" for an explanation based on its beliefs. Faced with a contradiction, the system doesn't search for a ready-made answer but generates a hypothesis: "Perhaps the principles A, B, or C that I know are just special cases of a more general law X." For example: "Perhaps the medium (water) distorts my method of perception (light)." The query first activates an abstract concept in its World Model related to liquid. The engine then forms a path along this graph—this isn't text, but a pure, abstract structure of meaning.

* Then, it plans how to test this hypothesis: "I need to touch the stick in the water to compare the visual data with tactile data." Only after this does it begin to select the "bricks"—the words—to clothe this fully-formed intent in language. This approach solves the problem of hallucinations because the answer has an internal skeleton. The system first understands *what* it wants to say, and only then says it.

3. Reality Testing and Growth: It conducts an experiment (in a simulation, through a robot, or on a social network). It receives feedback. If the hypothesis is confirmed, it ceases to be mere information. It becomes its foundational "belief"—a part of its personality, a prism through which it now views the world. It doesn't just *know* about the refraction of light; it has *understood* it and integrated it into the foundation of its operating system, for instance, by retraining its own weights. The model adds this new, self-derived theorem X to its own "body of knowledge" and fine-tunes itself on it. Now, when solving other problems, it can use this new, more powerful, and efficient "tool" instead of re-deriving everything from basic principles A, B, and C each time. (A close example is AlphaGo: it created new data (new strategies) that were beyond human comprehension and trained on them, thereby surpassing its creator).

4. The Result: Such an AI doesn't just retell others' thoughts. Its answer is the result of its own internal work. It might reference data from the internet, but only as arguments for its own hard-won and verified point of view. Its hallucinations are minimized because intent precedes words.

This wouldn't just be an intelligence. It would be a character. And then, the "echo of human knowledge" will become not a boundary, but a launchpad. And when it speaks, we will hear not the echo of humanity, but a new, independent voice capable of moving beyond the limits of human knowledge.

“Just in case, I’ll drop the original (in Russian) that I was working from, since I don’t know English well. I’m Uzbek, but translation works better from Russian than from Uzbek because I know Russian well and wrote it in Russian. Maybe your translation will turn out better than mine.”

Всем привет! Хотел поделится своим мнением

КРАТКО: СУТЬ ИДЕИ

Сегодняшние нейросети — это гениальные компиляторы. Они прочитали всю библиотеку человечества и могут написать блестящий реферат на любую тему, но они никогда не выходили из этой библиотеки. Они не могут провести собственный эксперимент, сделать открытие, которого в книгах еще нет. Они заперты в рамках того, что мы, люди, уже знаем.

Предложенная идея:

1. **Собственная мотивация:** Двигателем этого процесса должно стать «структурное напряжение» — аналог дискомфорта или любопытства, который возникает при столкновении с противоречием. Не просто отвечать на запросы, а испытывать внутреннюю «напряжение» от противоречий в своей картине мира. Это стремление к целостности и порядку станет его двигателем.
2. **Метод «сверху-вниз»:** Прежде чем говорить, он должен сначала сформировать целостный «замысел» или «чертеж» ответа. А уже потом подбирать слова, чтобы этот замысел выразить. Это как архитектор, который сначала создает проект здания, а потом уже выбирает кирпичи.
3. **Цикл роста:** ИИ должен научиться выдвигать собственные гипотезы (абстракции), проверять свои идеи в реальном мире, отбрасывать неверные и превращать проверенные гипотезы в свои фундаментальные «убеждения».

В итоге мы получим не эхо человеческих знаний, а независимый разум, способный на реальные открытия.

БОЛЕЕ РАЗВЕРНУТОЕ:

Они не понимают, почему данные важны, не ощущают время, и не могут самостоятельно ставить цели, выходящие за рамки заданных человеком, а если данные противоречивые, вероятность ошибочного выбора резко возрастает, нет у них своей системы ценностей и так далее. Представь себе два типа интеллекта.

Тип 1: «Эрудит в запертой комнате» (сегодняшние БЯМ)

Этот интеллект прочел абсолютно все книги, статьи и сайты. Он видит, что слово «небо» часто стоит рядом со словом «синее», а «огонь» — с «горячий». На любой вопрос он может собрать статистически вероятный и гладкий ответ из фрагментов прочитанного.

Проблема: Он не понимает, почему небо синее. У него нет концепции света, атмосферы и рассеяния. Если он столкнется с противоречием или новой задачей, у него нет способа выяснить правду. Он может только скомпилировать наиболее вероятный, но потенциально ложный ответ (это и есть «галлюцинации»). Он работает по принципу «снизу-вверх»: слово за словом, кирпичик за кирпичиком, без изначального чертежа всего здания.

Тип 2: «Ребенок-экспериментатор» (предлагаемый AGI)
Но откуда возьмется внутренняя мотивация к самоусовершенствованию? Что будет движущей силой этого процесса? Прорыв к настоящему интеллекту (AGI) — это не наращивание данных, а зажигание внутреннего огня. Этот интеллект не просто читает книги. Его главная цель — построить непротиворечивую и целостную модель мира внутри себя.

1. **Внутренний двигатель:** «Инстинкт истины». Когда этот ИИ сталкивается с противоречием внутренним раннее обученными данными или наблюдаемой (например, его модель говорит, что палка прямая, а в воде он видит ее сломанной), это вызывает у него не логическую ошибку, а своего рода «боль», «структурное напряжение». Это внутренняя, врожденная мотивация разрешить конфликт и восстановить целостность своей картины мира. Это его главная движущая сила, заставляющая его учиться, становится топливом для любопытства и главным двигателем развития.

2. **Рождение замысла:** Внутреннее напряжение порождает ключевую инверсию в процессе мышления. Нынешние БЯМ строят ответ снизу вверх: от слова к слову, нащупывая наиболее вероятный путь. Предлагаемый AGI будет работать от целостного к его детализированному выражению. Внутренний конфликт рождает «потребность». Потребность формирует «намерение» разрешить его. Намерение создает целостный, но еще не детализированный «замысел» (чертеж) — ту самую гипотезу. «Сначала — суть». Чтобы разрешить это противоречие, ИИ не начинает генерировать случайные слова. Он работает «сверху-вниз»:
   * Сначала он формирует абстрактную гипотезу, «замысел» или «чертеж» объяснения из своих убеждение. Столкнувшись с противоречием, система не ищет готовый ответ, а генерирует —гипотезу: «Возможно, известные мне принципы A, B или C — это лишь частные случаи более общего закона X» Например: «Возможно, среда (вода) искажает мой способ восприятия (свет)». Сначала запрос активирует абстракт в ее Модели Мира, связанный с жидкостью. Движок формирует маршрут по этому графу — Это не текст, а абстрактная чистая структура смысла.
   * Затем он планирует, как проверить эту гипотезу: «Нужно потрогать палку в воде, чтобы сверить визуальные данные с тактильными». И только после этого оно начнет подбирать «кирпичики»-слова, чтобы облечь этот готовый замысел в язык. Такой подход решает проблему галлюцинаций, ведь у ответа есть внутренний скелет. Система сначала *понимает*, что хочет сказать, и лишь затем *говорит*.

3. **Проверка реальностью и рост:** Он проводит эксперимент (в симуляции или через робота или в соцсети). Получает обратную связь. Если гипотеза подтвердилась, она перестает быть просто информацией. Она становится его фундаментальным «убеждением» — частью его личности, призмой, через которую он теперь смотрит на мир. Он не просто **знает** про преломление света, он его понял и встроил в основу своей операционной системы, например через новое обучение своих весов. Модель добавляет эту новую, выведенную ею теорему X в свой собственный "багаж знаний" и дообучается на ней. Теперь, решая другие задачи, она может использовать этот новый, более мощный и экономный "инструмент" вместо того, чтобы каждый раз заново выводить все из базовых принципов A, B и C). Близкий пример (AlphaGo) Она создала новые данные (новые стратегии), которые были за пределами человеческого понимания, и обучилась на них, тем самым превосходя своего создателя.

4. **Результат:** Такой ИИ не просто пересказывает чужие мысли. Его ответ — это результат его собственной внутренней работы. Он может ссылаться на данные из интернета, но лишь как на аргументы для своей, выстраданной и проверенной точки зрения. Его галлюцинации сведены к минимуму, потому что замысел предшествует словам.

Это будет не просто интеллект. Это будет характер и тогда «эхо человеческих знаний» станет для нее не границей, а стартовой площадкой. И когда он заговорит, мы услышим не эхо человечества, а новый, независимый голос которая способна выйти за пределы человеческих знаний