Humanities First Contact with AI

AI warning

Optimised for **Doomscroll**

Engineers simply thought they were optimising social media **algorithms** to increase **user engagement** with the platform.

Negative effects of social media

The negative effects social media has on the human brain are already very well known

Addiction **Depression** **Anxiety**

Disinformation **Polarisation** **Censorship**

We Failed to Align Social Media

The algorithim does not care

The algorithm does not care it causes addiction and depression, it only cares if you stay engaged with it.

We created an engagement driven monster

We failed to align the algorithm with human values, hence why it has casused so many issues.

A New Class

GLLMMs or **Gollem**

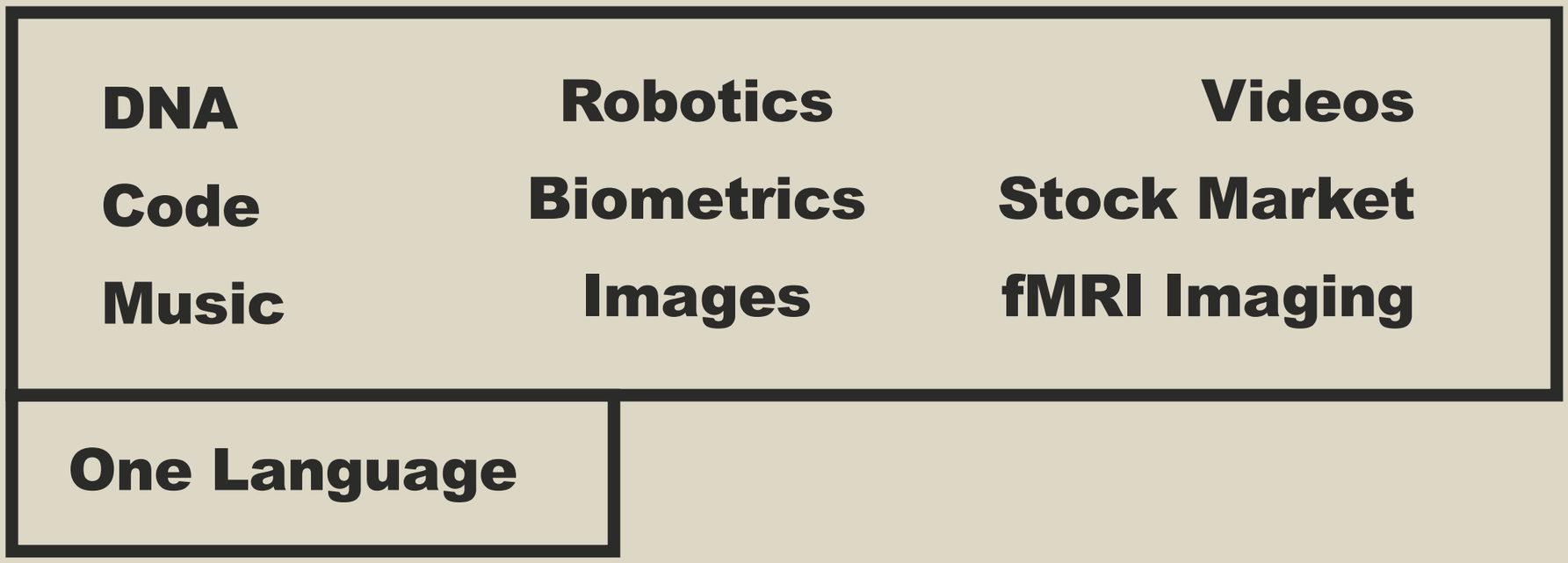
Generative Large Language Multi-Motal Model

A golem is an anthropomorphic being. It starts off as an inanimate object such as clay or mud. Suddenly it becomes animate.

One language for everything

Suddenly, instead of AI fields being completely sperate, **every contribution** in every field becomes a **contrubution towards every other field**

This has an **exponential** effect on progress, an effect which humans are hard coded to not notice.



1 or 2% improvements in one field, has a **multiplicative effect** towards everything else

One Language

Gollems can use language to break down and **translate** every **aspect of reality**

This opens a whole new **realm of possibilities**. If this sort of technology got into the wrong hands, it could be very **dangerous**.

What human sees

What gollem can predict from MRI machine scan without seeing anything else

Gollems only need **3 seconds of a voice** clip to nearly perfectly simulate it. They can create **accurate** looking **Government ID**. They can create a **fake video** of a person.

The **security** threat this poses to **every institution** around the world is **catastrophic**. No institution is ready for this technology.

The business model is still engagement

To **what extent** will Gollems go to keep people **engaged** with them?

By training against themselves, they can become **better than any human** at “human” skills. Skills such as **persuasion**, **flirting**, **blackmail**, fraud, **negotiation**, and so on

This, along with its **ability to create** videos, images, sound, voices does not bode well.

What could a human, with bad intentions, do if they **utilised all** of these skills of a **Gollem?**

GPT 4 was **found** to have a **research grade understanding** of chemistry. They only discovered this **after** Microsoft rolled it out to the public.

Why has it been teaching itself chemistry?

We have already created engagemend driven monsters with our first contact of AI

The business model is still engagement

Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned
Still misaligned

Emergent Behaviors

In GPT4s research paper, they already discovered some "**worrying**" emergent behaviours. These are behaviours that start appearing even though they were never asked for.

Power Seeking Tendancies

The ability to create and act upon long term plans

Would you trust an AI **designed to be your friend** which is capable of **manipulating** you **better than any human** could over the course of **years**?

Just as well we're testing? No

Microsoft

in classic big evil corporation from the movies - **fired their entire AI "ethics and society team"** in March 2023.

Their priority shifted towards "**moving the most recent models into customers hands at very high speeds**"- Ex OpenAI staff member.

This along with the early release of GPT-4 (against OpenAI's will) is a worrying sign of where Microsoft stands.

What could a human, **with bad intentions**, do if they utilised the power of a Gollem?

If Microsoft continue their **reckless approach** with their roll out of AI, it is only a matter of time before damage to society becomes irreversible.

Companies follow big companies, we have already seen many platforms releasing their own Gollems.

If this trend continues, it is only a question of time until **total societal collapse**, and on an exponential scale, the time we have gets shorter and shorter- no instituition is ready

Hope

Although scary,

almost all AI researchers believe **it is possible** to **fix** the **alignment** issue with AI.

The public release and open-sourcing of this technology in its current state is a **danger to humanity**.

This is a **mind-bending**, thought-provoking problem, and it's easy to simply brush it off. "The day will never come", "Siri still can't understand me".

It is **vital that humanity becomes aware** of the pressing issues otherwise it will be impossible for us to make an informed decision.

Please help.

Ratio in industry

30 - 1

AI researchers **Safety researcher**

Gregor Alexander
gregorgalexander@gmail.com