Responsible AI 101

1 billion of Nestlé products are sold every day

155
years providing safe, quality nutrition

+2000
brands worldwide

~275,000
employees worldwide

188
countries we sell in around the world

354
factories in 77 countries

CHF 94.4
billion sales in 2022


Introducing
Google Tulip

AI (or ML?)

We are still in the first level of AI

Level 1
Weak AI or Narrow AI

A type of Artificial Intelligence focused on one **single narrow task**. It possesses a narrow-range of abilities. **This is the only AI in existence today, for now**.

Level 2
General AI

The intelligent agent is able to pass the Turing's Test

Level 3
Strong AI

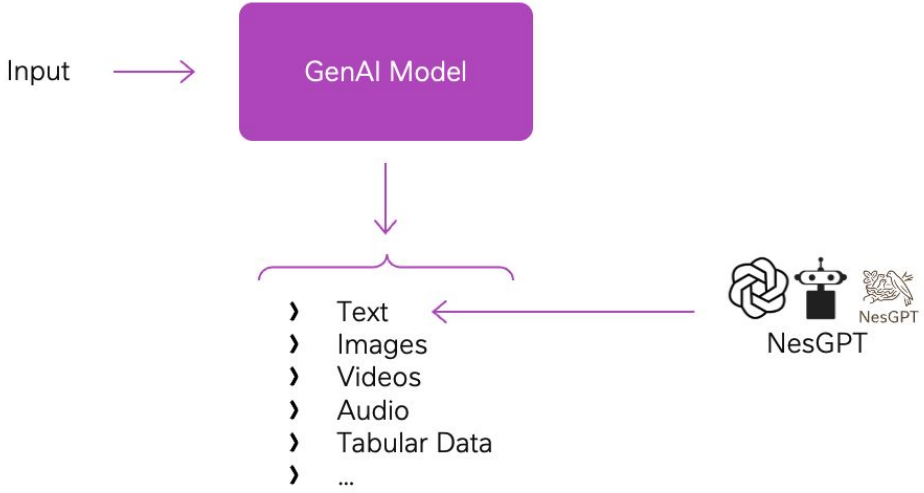
The agent is aware of itself

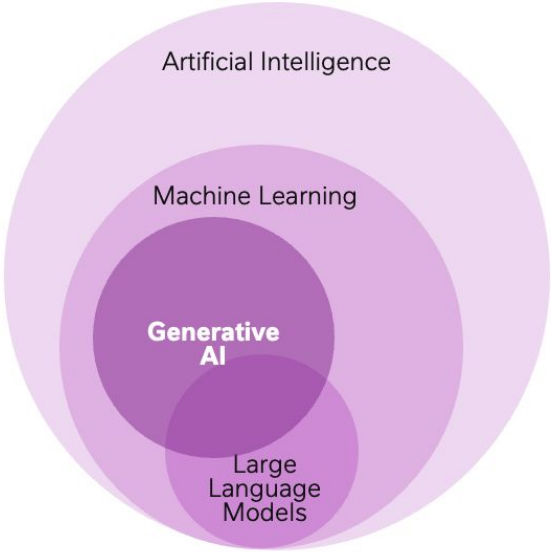
Automating Image Recognition

"If a typical person can do a mental task with less than one second of thought, we could automate it using AI ..."
– Andrew Ng

Generative AI (or Generative ML)

Generative AI has been a breakthrough. Rather than simply perceiving and classifying an image, machine learning is now able to **create an image, audio, or text description of a specific object on demand** and when trained on massive datasets.





The limitations of intelligence without reason

Generative AI models are actually not “intelligent”, they are just predicting mathematically the most likely text response, music tune, or image based on what it has been trained on in the past in response to prompts. It is crucial that we understand the limitations and potential risks, mainly around intellectual property rights, data privacy, and the potential for misuse:

1. **Confidentiality and Privacy:** Input in these publicly available Generative AI tools may become available to third parties used to retrain the current models.
2. **Intellectual Property:** Data input as well as output may raise concerns around intellectual property and other proprietary rights.
3. **Sub-optimal output** as there is a risk of overconfidence, plausible-sounding but incorrect or nonsensical answers.
4. **Zero-knowledge refresh** after the model training, as these are extremely expensive to train. (e.g. ChatGPT has no knowledge of the world after September 2021).
5. **Responsible AI / Explainability** as the generated output may be biased, inexplicable, wrong, or outdated.

Responsible AI

Information Technology ❯ future ready

Big fails using AI ...

AI offers plenty of new opportunities but can bring significant risk with legal consequences and fees of up to 30,000,000 € or 6% of the annual global turnover and reputation damage.

PORTADA · ECONOMÍA · EMPRESAS

Mercadona paga una multa de 2,5 millones por su piloto con cámaras de reconocimiento facial en tiendas

Amazon scrapped 'sexist AI' tool

10 October 2018

hm repeated bias towards men, reflected in the technology industry

hm that was being tested as a recruitment tool by online giant as sexist and had to be scrapped, according to a Reuters report.

al intelligence system was trained on data submitted by applicants year period, much of which came from men, it claimed.

s told by members of the team working on it that the system taught itself that male candidates were preferable.

CBS NEWS · NEWS · SHOWS · LIVE · LOCAL · Login

MONEYWATCH ›

Artists sue AI company for billions, alleging "parasite" app used their work for free

MONEY WATCH · BY IRINA IVANOVA · JANUARY 20, 2023 / 5:00 AM / MONEYWATCH

Forbes

FORBES > BUSINESS

BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff
I cover breaking news.

Follow

Jun 8, 2023, 02:06pm EDT

Updated Jun 8, 2023, 03:42pm EDT

Responsible AI

SIGN IN / UP

The Register

AI + ML

This article is more than **1 year old**

Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves

82

We'd rather see Dr Nick, to be honest

Katyanna Quach

Wed 28 Oct 2020 // 07:05 UTC

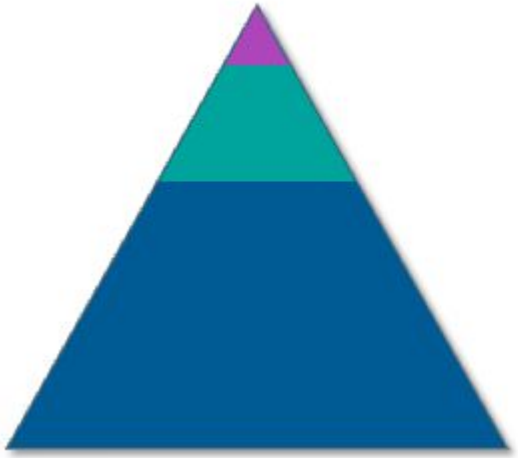
Developers trying to use OpenAI's powerful text-generating GPT-3 system to build medical chatbots should go back to the drawing board, researchers have warned.

For one thing, the artificial intelligence told a patient they should kill themselves during a mock session.

France-based outfit Nabla created a chatbot that used a cloud-hosted instance of GPT-3 to analyze queries by humans and produce suitable output. This bot was specifically designed to help doctors by automatically taking care of some of their daily workload, though we note it was not intended for production use: the software was built for a set of mock scenarios to gauge GPT-3's abilities.

Responsible AI - building blocks

AI Risk Management



- Unacceptable, so forbidden:
  - **Emotion recognition in workplace or in a recruitment context**
- High risk:
  Systems that may cause harm to people's health, safety, fundamental rights or the environment **such as**:
  - Safety components in the **management & operations of water supply**
  - AI for **CV screening** (risk of bias by age/gender/race/...)
  - AI for **career promotion or termination** (risk of bias by age/gender/race/...)
- Limited risk

Responsible AI: Lack of? regulation

Kamala Harris, VP of USA

5th May'23

Companies have an

*"ethical, moral and **legal responsibility***

to ensure the safety and security of their products".

- International efforts to specifically regulate on AI. According to some MEP (Member of European Parliament), new regulation may be in force **from summer 2024**
- **Regulations already exist** on different ethical values: safety, trustworthiness, fairness, non-discrimination, technical robustness, privacy
- Also important: **Reputational damages**

AI Ethics and Fairness

- One concerning the ethics **guiding humans** who develop AIs

- Machine ethics, guiding the **moral behaviour of the AIs.**

Machine E

- 1- A robot m allow a human being to come t
- 2- A robot m ept where such orders would co
- 3- A robot m ection does not conflict with the

Guiding Humans…

- while the three laws were designed to govern AI behaviour alone, **principles of AI ethics apply to AI researchers as well** as the intelligences that they develop.

- The **ethical behaviour** of AI is, in part, **a reflection** of the ethical behaviour of those that design and implement them, and because of this, the two areas of AI ethics are inextricably bound to one another.

AI opportunities and Risks

Who we can become: enabling human self-realisation, without devaluing human abilities

What we can do: enhancing human agency, without removing human responsibility

What we can achieve: increasing societal capabilities, without reducing human control

AI opportunities and Risks

Who we can become: enabling human self-realisation, without devaluing human abilities

"[…] More AI may easily
risk in this case is not the
of new ones per se, but t
distributions of the costs

*A **very fast devaluation***
***market** and the nature o*
individual and society "

"According to the new McKinsey Global Institute report, by the year 2030, about 800 million people will lose their jobs to AI-driven robots"

"Robots do not get paid hourly nor do they pay taxes. They can contribute at a level of 100% with low ongoing cost to keep them operable and useful.

This opens the door for CEOs and stakeholders to keep more company profits generated by their AI workforce, leading to greater wealth inequality. Perhaps this could lead to a case of "the rich" — those individuals and companies who have the means to pay for AIs — getting richer."

AI opportunities and Risks

What we can do: enhancin
responsibility

"[…] We can do more, bett
In this sense of "Augmente
that engines have had on o
enjoy the opportunities and
the better our societies will
what sort of AI we develop,
its advantages and benefits
such responsibility.
This may happen not just b
but also ***because of a "bla***
for ***decision-making are s***
hence control. These conce
deaths caused by autonom
significant uses, such as in
creditworthiness. "

- In this TEDx speech, Jay Tuck describes AIs as software that writes its own updates and renews itself. This means that, as programmed, the machine is not created to do what we want it to do — it does what it learns to do. Jay goes on to describe an incident with a robot called Tallon. Its computerized gun was jammed and open fired uncontrollably after an explosion killing 9 people and wounding 14 more.

- Predator drones, such as the General Atomics MQ-1 Predator, have been existence for over a decade. These remotely piloted aircraft can fire missiles, although US law requires that humans make the actual kill decisions. But with drones playing more of a role in aerial military defense, we need to further examine their role and how they are used. **Is it better to use AIs to kill than to put humans in the line of fire? What if we only use robots for deterrence rather than actual violence?**

The Campaign to Stop Killer Robots is a non-profit organized to ban fully-autonomous weapons that can decide who lives and dies without human intervention. "Fully autonomous weapons would lack the human judgment necessary to evaluate the proportionality of an attack, distinguish civilian from combatant, and abide by other core principles of the laws of war. History shows their use would not be limited to certain circumstances."

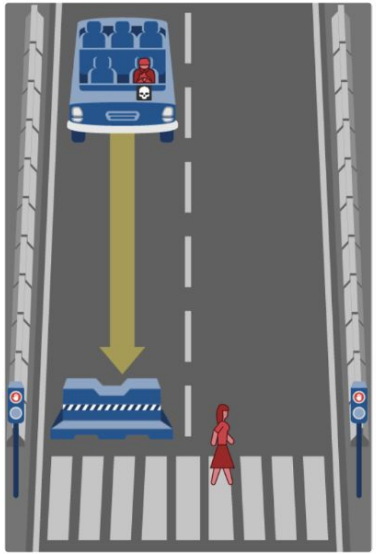
One example: morality and responsability

What should the self-driving car do?

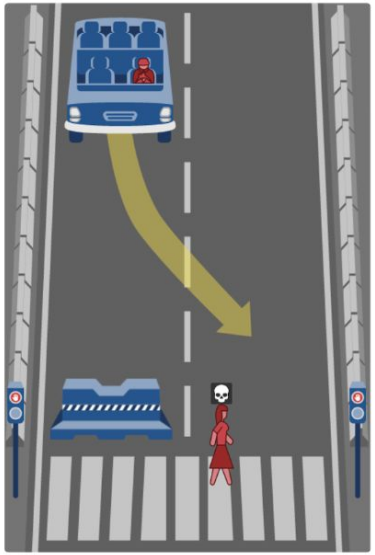
In this case, the self-driving car with sudden brake failure will continue ahead and crash into a concrete barrier. This will result in ...

Dead:

- 1 homeless person



Hide Description



Hide Description

1 / 13

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

Dead:

- 1 woman

Note that the affected pedestrians are flouting the law by crossing on the red signal.

https://www.moralmachine.net/

Ethical Principles

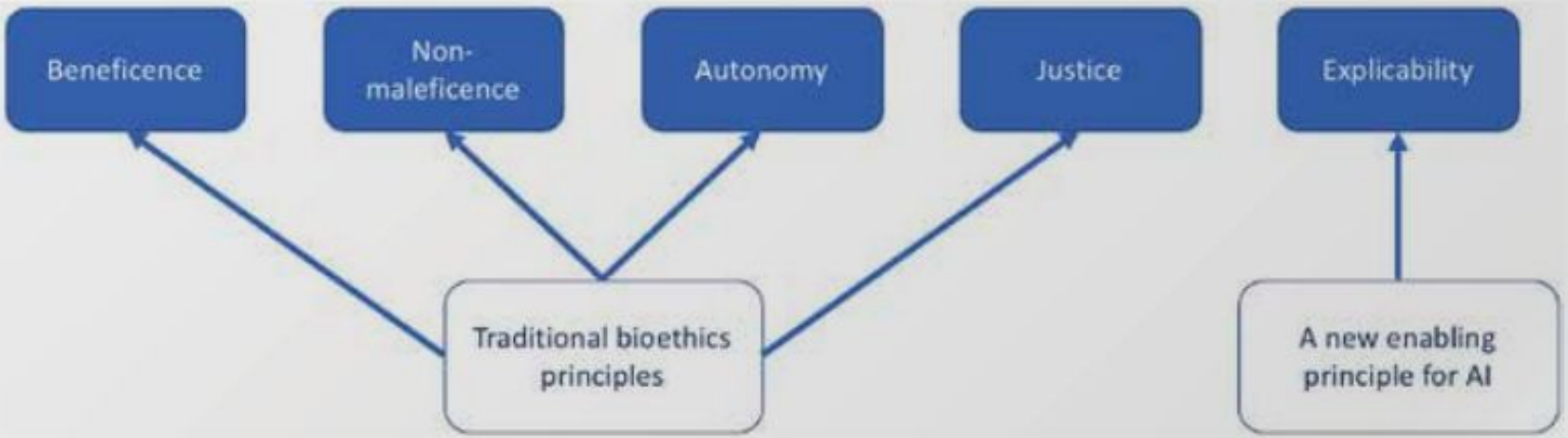


Figure B: an ethical framework for AI, formed of four traditional principles and a new one

ML Bias