DIGIPHILE

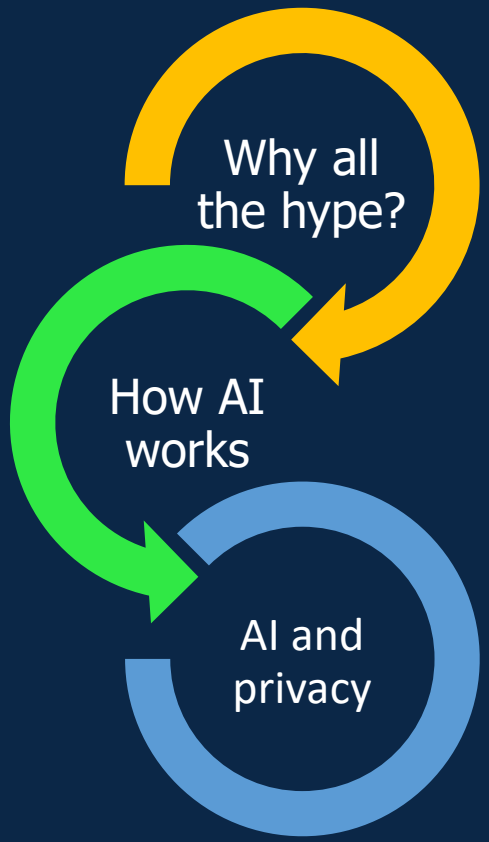
Machines behaving badly?

Phil Lee, Managing Director

DIGIPHILE

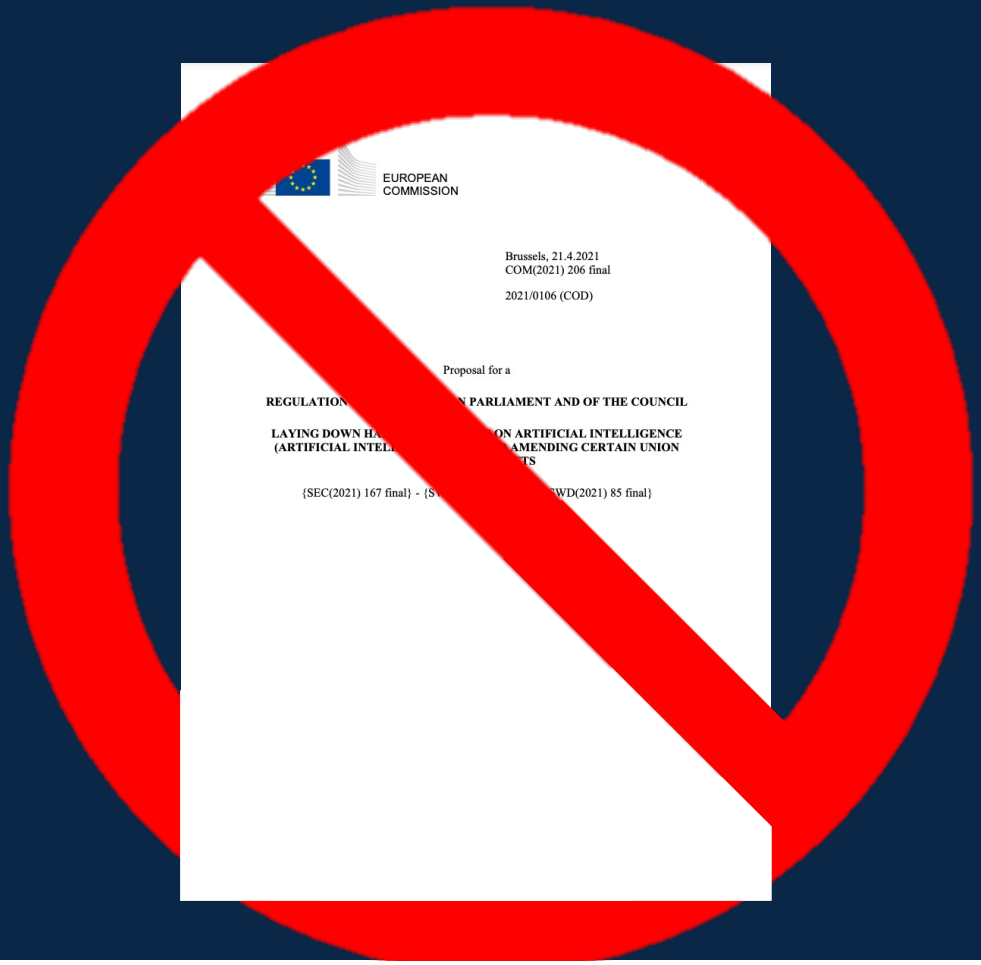
Our agenda

- Why all the hype
  - What is AI?
  - AI making headlines
  - How does AI work?
- How AI works
  - AI, maths and data
  - The three main types of AI
  - Neural nets and deep learning
- AI and privacy
  - Framing AI concepts under the GDPR
  - Key privacy challenges
  - Integrating AI into products



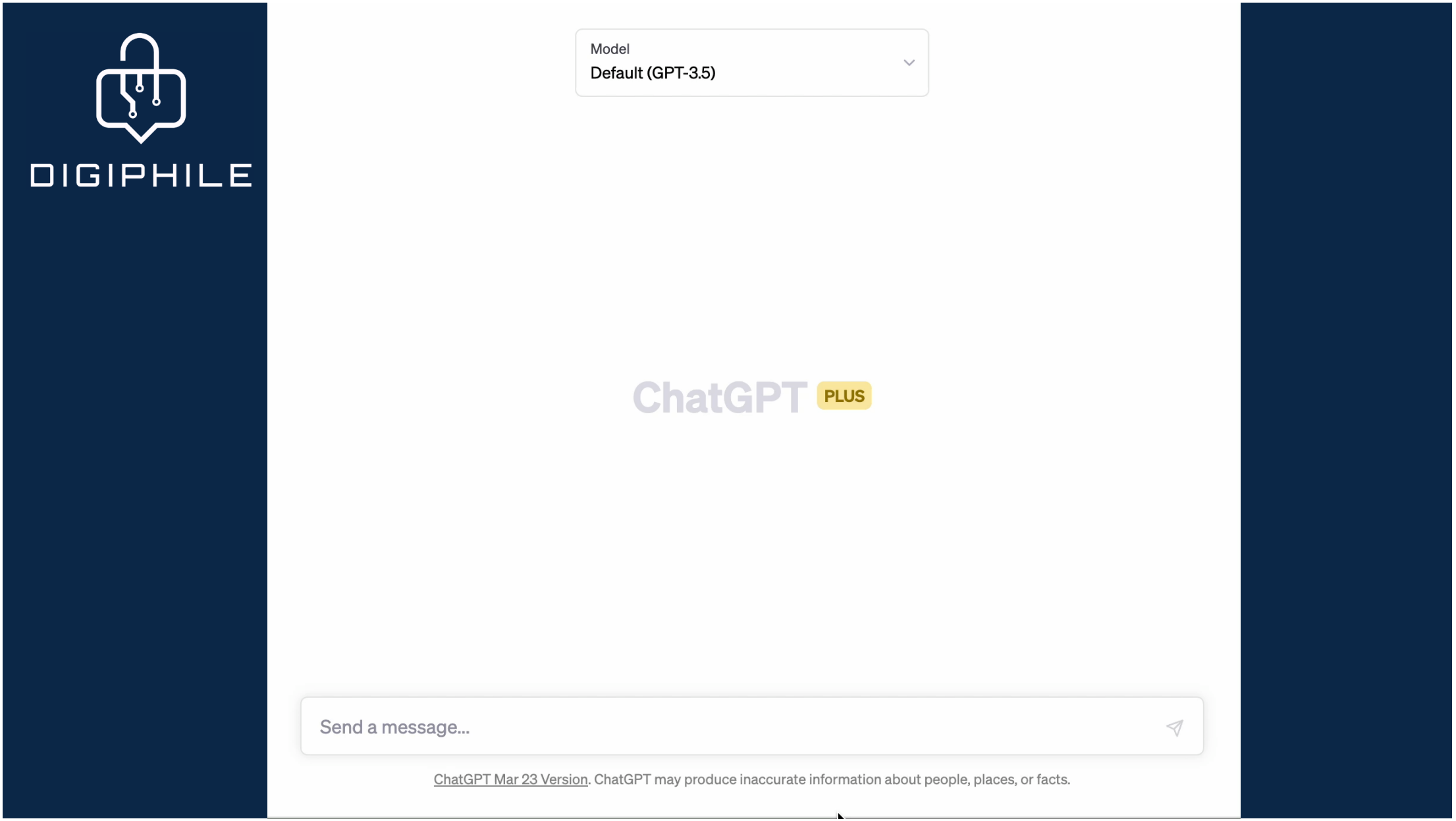
DIGIPHILE

What we’re not discussing (today)



DIGIPHILE

Part 1: Why all the hype?


DIGIPHILE
Model
Default (GPT-3.5)
ChatGPT PLUS
Send a message...
ChatGPT Mar 23 Version. ChatGPT may produce inaccurate information about people, places, or facts.

DIGIPHILE

What is AI?

SJ

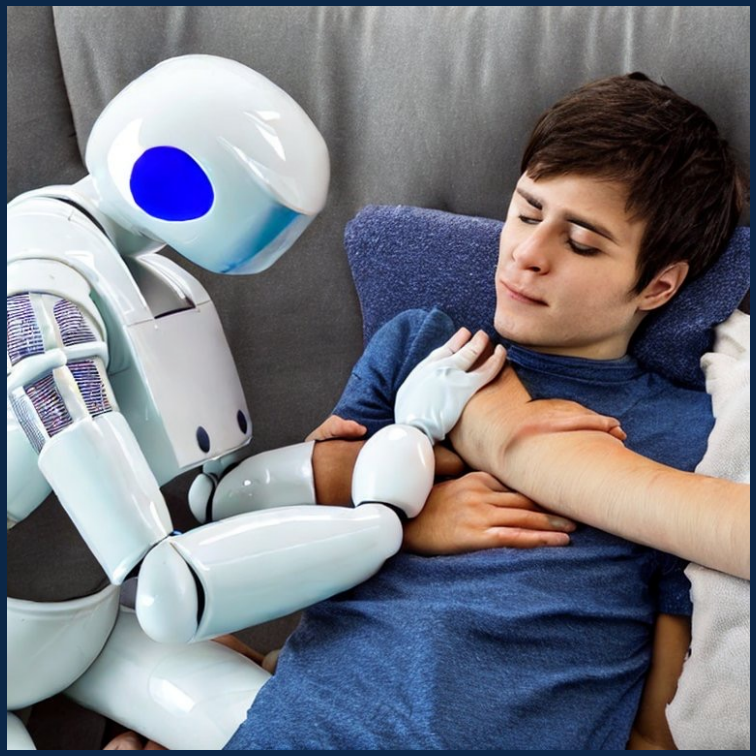
what is artificial intelligence

Artificial intelligence (AI) is the simulation of human intelligence in machines that are programmed to perform tasks that typically require human intelligence, such as visual perception, speech recognition, decision-making, and natural language processing. AI is a broad field that includes various subfields such as machine learning, natural language processing, computer vision, robotics, and expert systems. AI systems can be designed to work in a variety of environments, including healthcare, finance, transportation, manufacturing, and many others. AI has the potential to revolutionize many industries and improve our daily lives, but it also raises ethical, social, and economic concerns that need to be addressed.

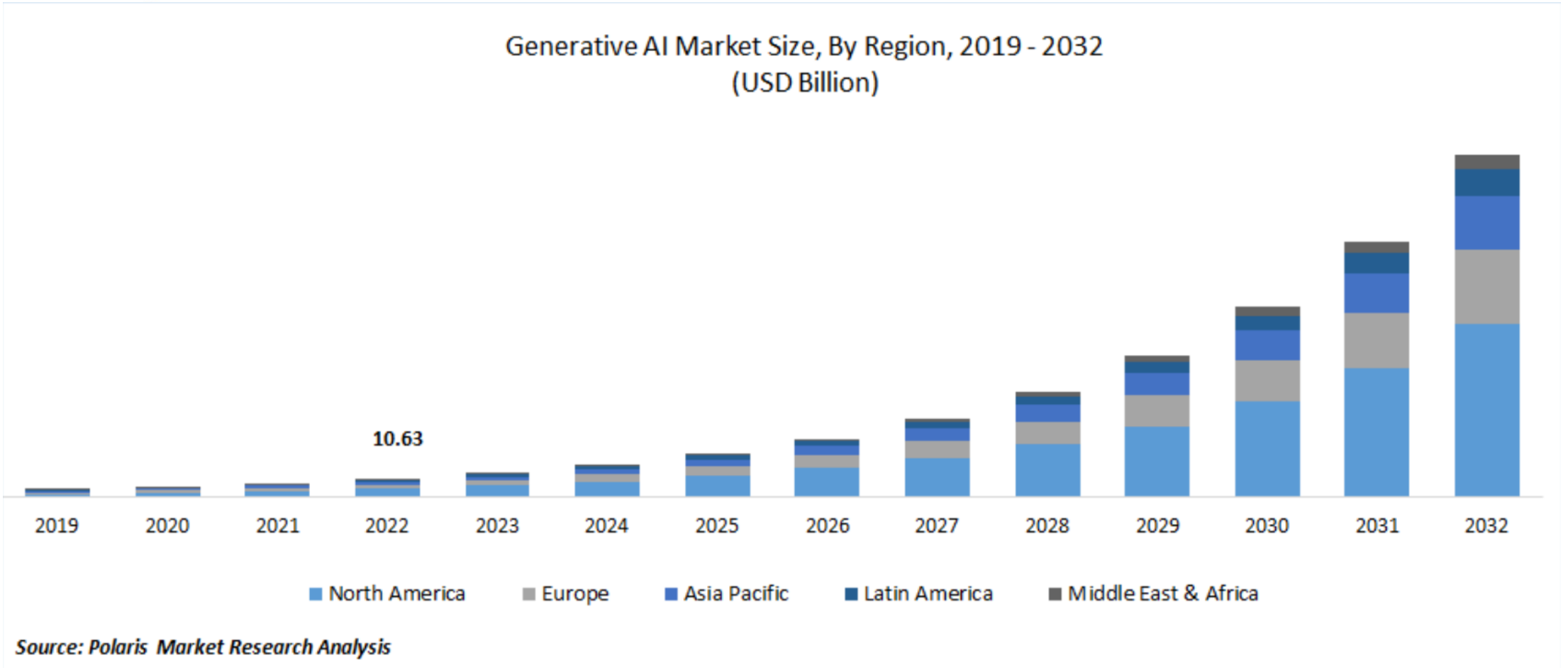
DIGIPHILE

Is AI something to fear?

Or something to cheer?



Both images created with Stable Diffusion



Source: https://www.polarismarketresearch.com/industry-analysis/generative-ai-market

Artificial Intelligence APIs Landscape - January 2023

Eden AI

Source: https://www.edenai.co/post/a-comprehensive-view-of-all-ai-apis-in-2023?referral=ocr-how-to

DIGIPHILE

AI in the news

BBC Sign in Home News Sport Weather iPlayer

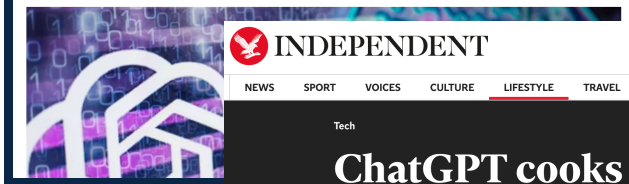
NEWS

Home | Cost of Living | War in Ukraine | Climate | UK | World | Business | Politics | Culture | Tech

Technology

ChatGPT banned in Italy over privacy concerns

1 April



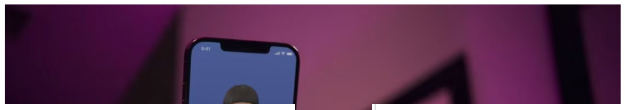
REUTERS® World Business Markets Legal Breakingviews Technology Investigations

Technology

2 minute read · February 3, 2023 9:37 PM GMT · Last Updated 3 months ago

Italy bans U.S.-based AI chatbot Replika from using personal data

By Elvira Pollina and Martin Coulter



euronews.next BIZTECH NEWS MONEY WORK MOBILITY HEALTH HOME OPINION SERIES

Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

INDEPENDENT

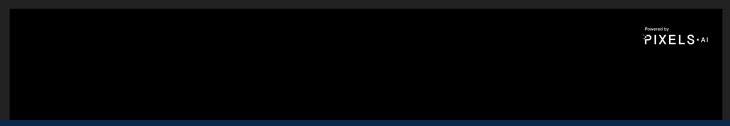
NEWS SPORT VOICES CULTURE LIFESTYLE TRAVEL PREMIUM MORE

Tech

ChatGPT cooks up fake sexual harassment scandal and names real law professor as accused

'When first contacted, I found the accusation comical. After some reflection, it took on a more menacing meaning'

Vishwam Sankaran • Thursday 06 • 3 Comments

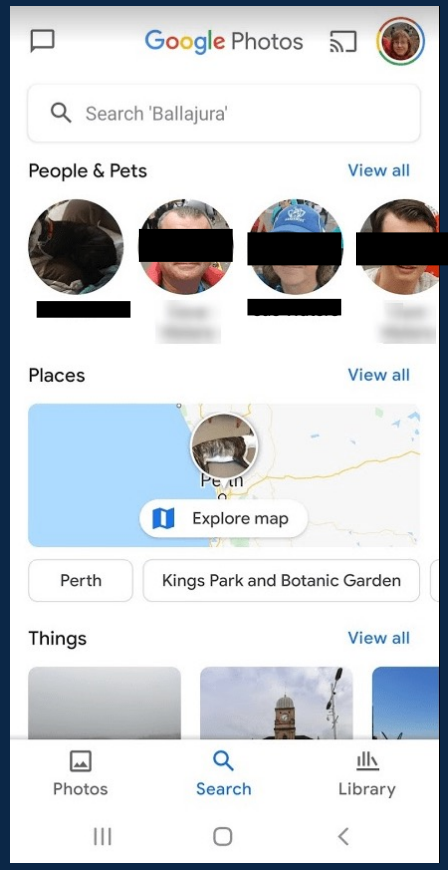


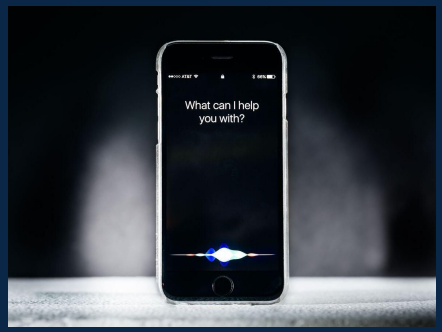
Tesla driver killed while using autopilot was watching Harry Potter, witness says

Driver in first known fatal self-driving car crash was also driving so fast that 'he went so fast through my trailer I didn't see him', the truck driver involved said

DIGIPHILE

But it’s not all bad





DIGIPHILE

But it’s not all bad (2)

DIGIPHILE

Part 2: How AI works

DIGIPHILE

AI is powered by algorithms

- $f(a) = 2a + 31$
- $f(a, b, c) = (a + b)^2 / c + 24$
- $f(a, b, c \ldots z) =$ 
- Now imagine an algorithm with millions of parameters*

* Reportedly 1 trillion in ChatGPT4

DIGIPHILE

AI is also powered by data

- Complete works of Shakespeare = around 5Mb
- Encyclopedia Britannica = around 1Gb (1Gb = 1024Mb)
- Human genome = around 3Gb
- ChatGPT 3 training data = 570Gb (570 Encyclopedia Britannicas!)

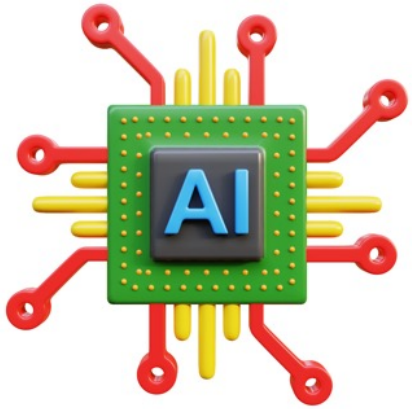
DIGIPHILE

Training AI (1): Supervised learning

Label: Cat

Label: Dog

Training

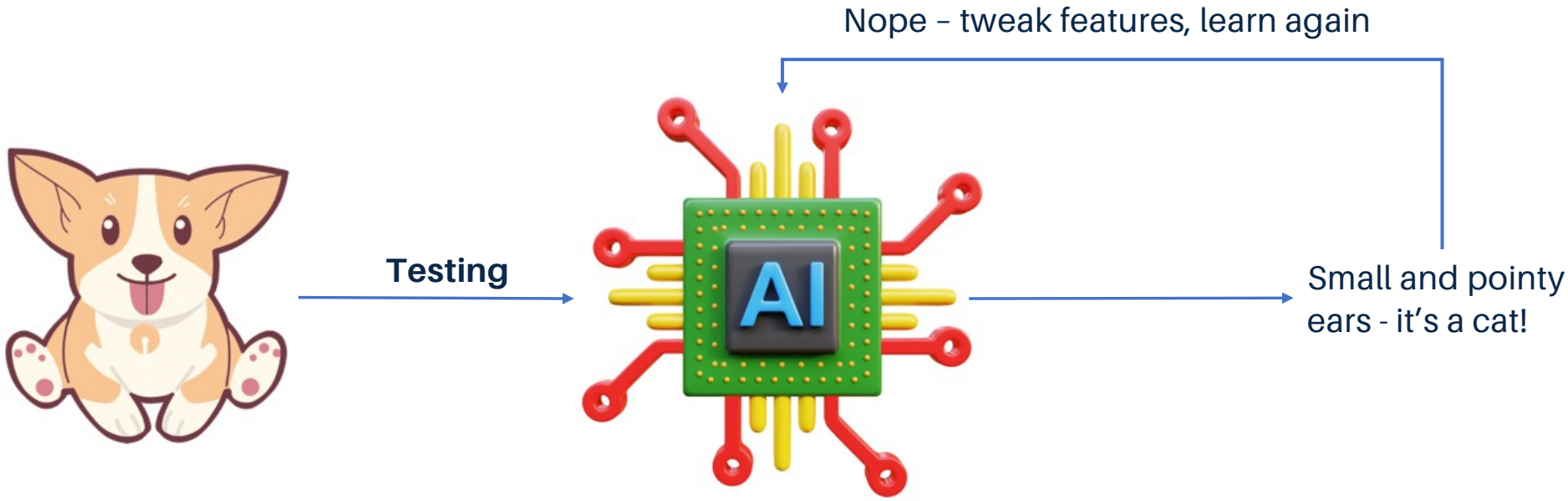


Cat features
Sharp claws
Whiskers
Pointy ears

Dog features
Bigger
Long tongue
General dog shape

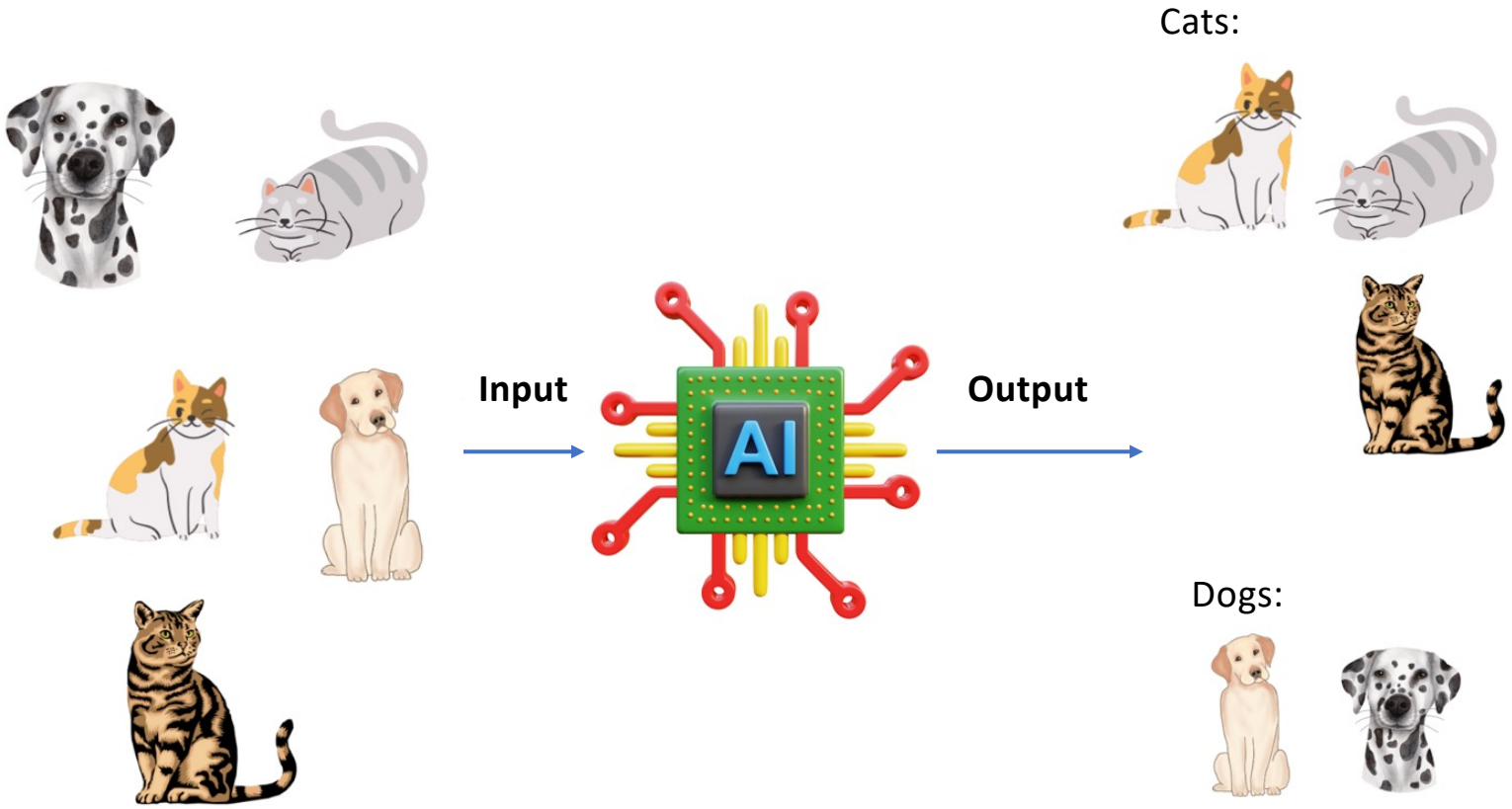
DIGIPHILE

Training AI (1): Supervised learning (2)



DIGIPHILE

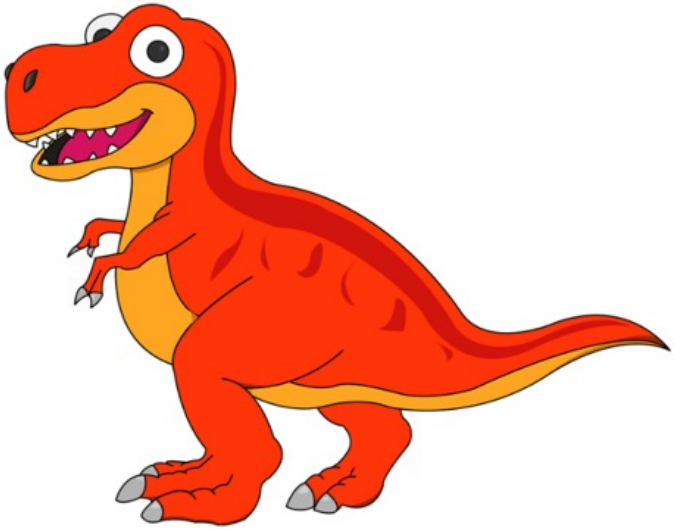
Training AI (1): Supervised learning (3)



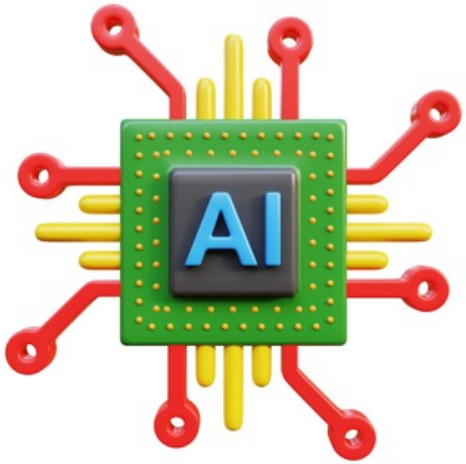
DIGIPHILE

Training AI (1): Supervised learning (4)

But be careful…



Input

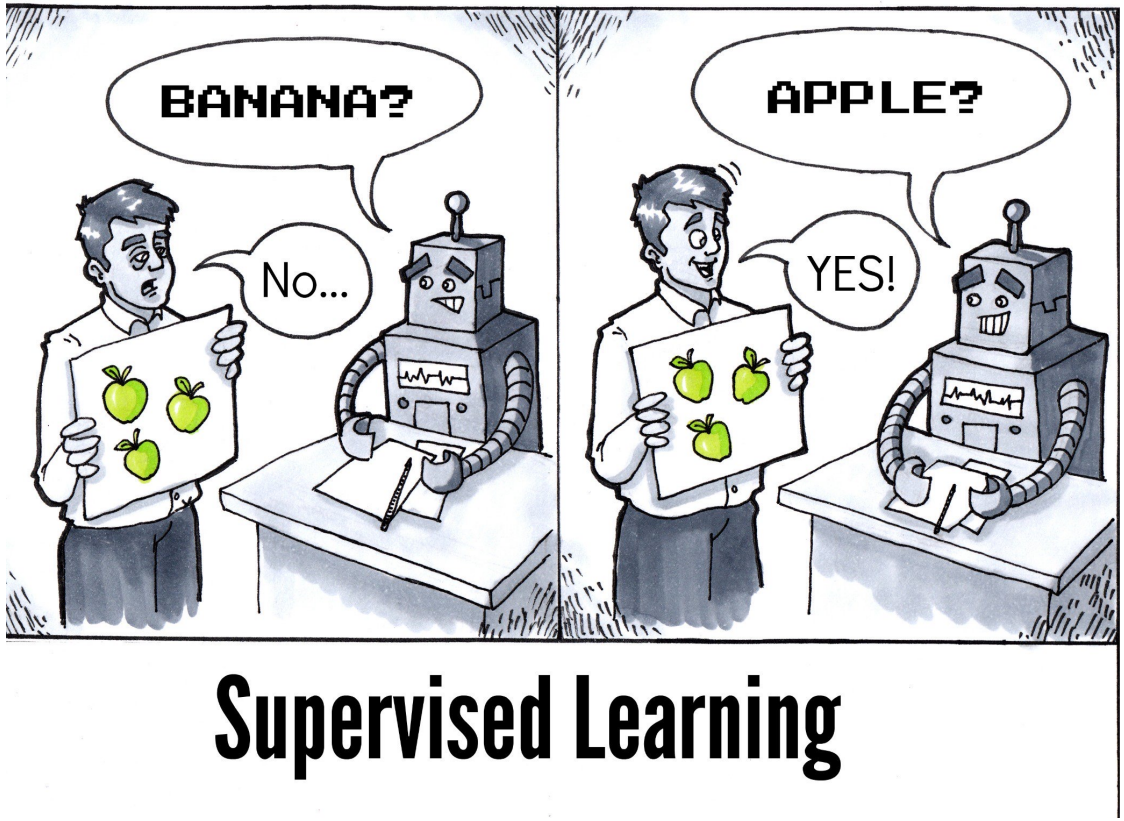


Output

It's big – must be a dog!

DIGIPHILE

Training AI (2): Unsupervised learning

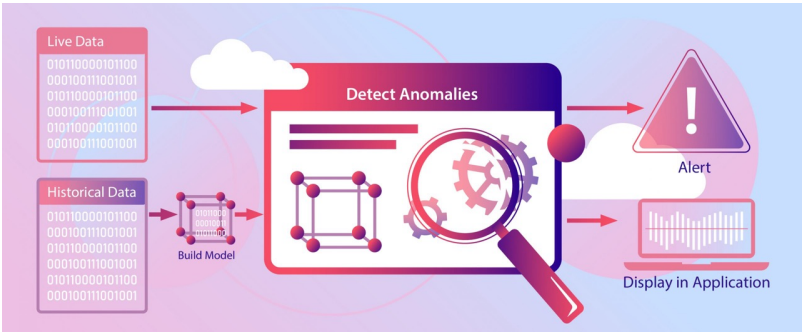


Source: https://twitter.com/athena_schools/status/1063013435779223553 (sharing an image by @Ciaraioch)

DIGIPHILE

Training AI (2): Unsupervised learning

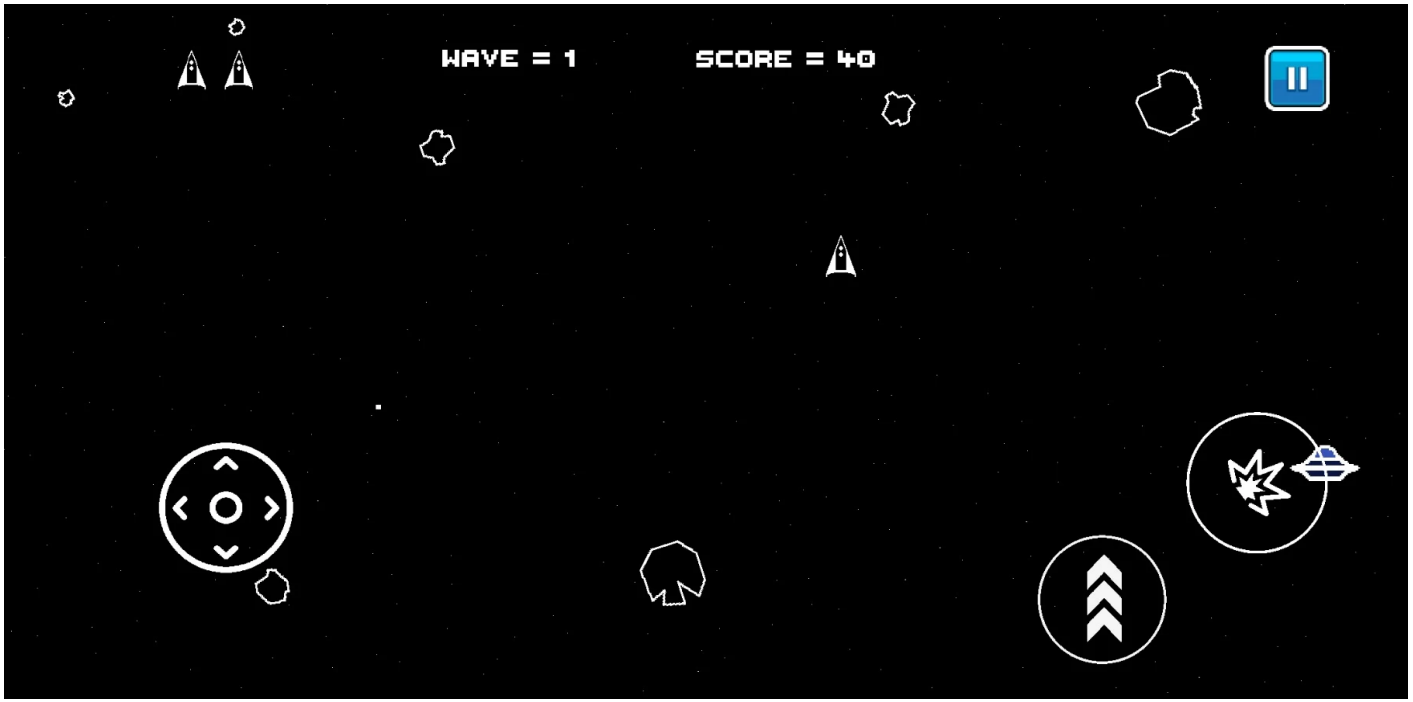
- No labels / no training data
- Groups naturally occurring patterns in data
- Used for clustering + anomaly detection
- E.g. fraud detection, network intrusion detection



Source: https://perfectial.com/blog/fraud-detection-machine-learning/

DIGIPHILE

Training AI (3): Reinforcement learning



- Reward-based learning
- E.g. Asteroids game

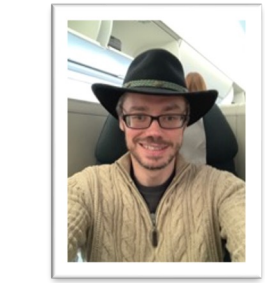
  Shoot asteroid = +10
  Get hit = -200
  Spend bullets = -1
  Spend fuel = -5

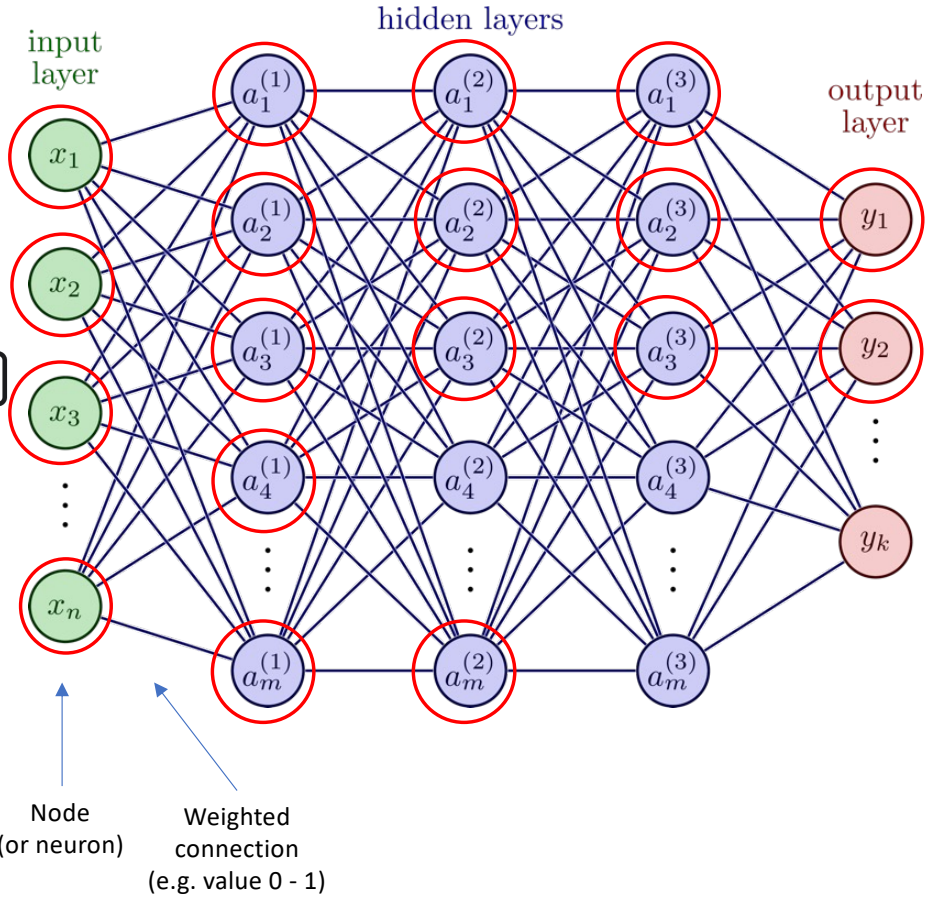
- Used (e.g.) in autonomous driving, teaching robots to walk

Source: Asteroids: Space Defense (MM Retro Games) available on Google Play:
https://play.google.com/store/apps/details?id=com.MMRetroGames.AsteroidsSpaceDefense

DIGIPHILE

The technology behind (most) AI: Deep learning and neural nets

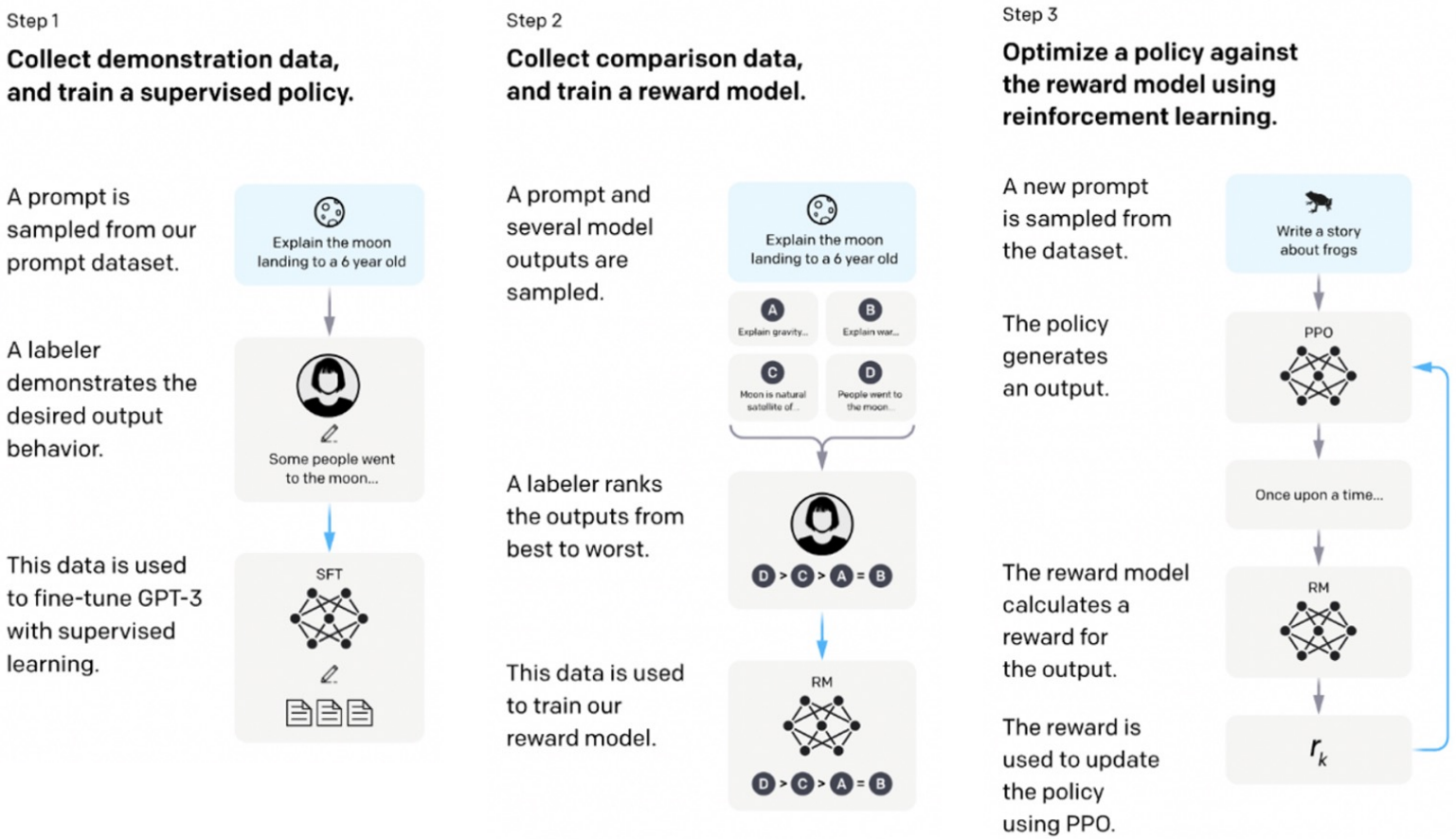




When y_1 or y_2 fire, it's Phil!

DIGIPHILE

So how does ChatGPT work?



DIGIPHILE

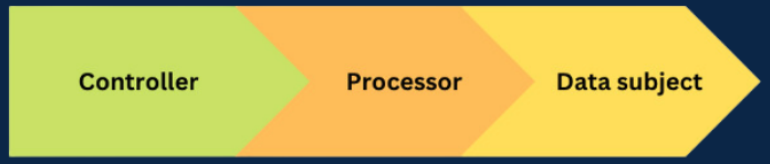
Part 3: AI and privacy

DIGIPHILE

Framing AI concepts under the GDPR

- Your role determines your GDPR responsibilities…*

- …but how to AI roles map to GDPR roles?



- …and the role you play will depend on the specific data!

* Note that EU's proposed AI Act will also introduce roles of "importer", "distributor" and "product manufacturer"

DIGIPHILE

Framing AI concepts under the GDPR

Data type	What's it used for?
Training data	Data used to **train** the AI model (e.g. the labelled data in supervised learning)
Testing data	Data used to **test** the AI model is functioning correctly (which may be a subset of the training data)
Input data	Data **entered into the AI model** to generate an output (e.g. instructions to ChatGPT to: "Write a privacy rock song in the style of Meatloaf")
Output data	Data **returned by the AI model** in response to its input (e.g. The image output in response to instructions to "Create a picture of King Charles III in the style of Leonardo da Vinci")
Any other data?	For example, login credentials to access the AI platform, telemetry data about use of the AI platform etc.?

DIGIPHILE

Assessing GDPR roles - example 1

AI - B2C controller/processor relationships (typical example)

AI data type	Consumer	B2C service	AI provider
Input data	Data subject	Controller	Processor *
Output data	Controller (Household exemption?)	Controller	Processor *
Training data	N/A	N/A	Controller

* Assumes input/output data not used for training

DIGIPHILE

Assessing GDPR roles - example 2

AI - B2B controller/processor relationships (typical example)

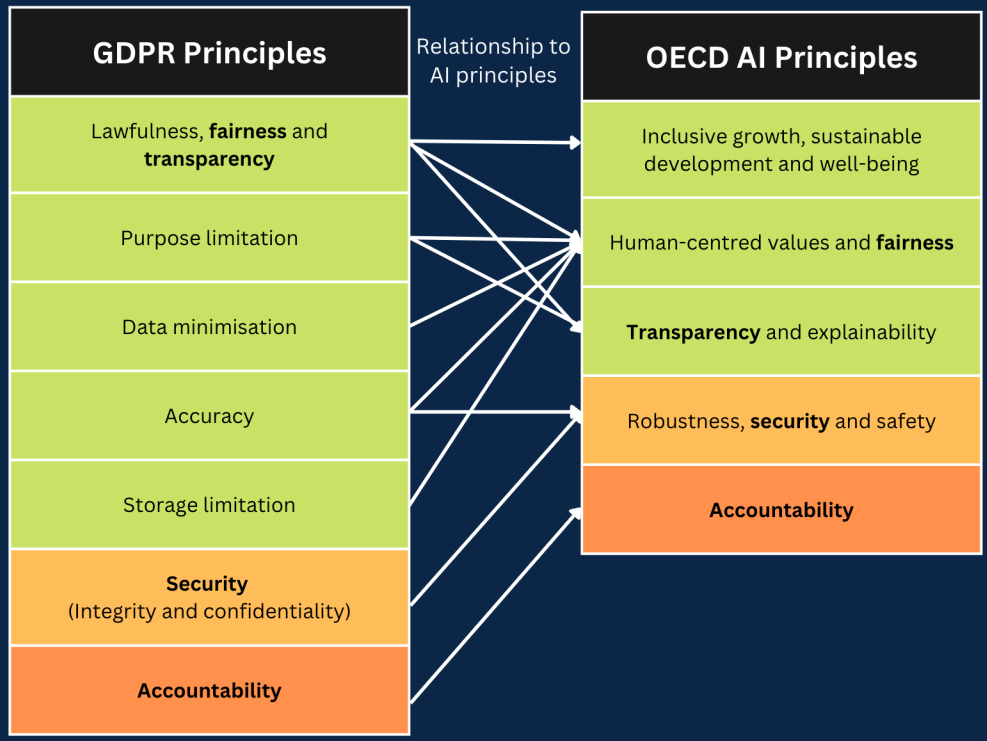
AI data type	Enterprise customer	B2B service	AI provider
Input data	Controller	Processor	Processor*
Output data	Controller	Processor	Processor*
Training data	N/A	N/A	Controller

* Assumes input/output data not used for training

DIGIPHILE

Framing AI concepts under the GDPR

- International community coalescing around certain key principles for trustworthy AI
- Examples include OECD principles, G20 principles etc.
- …but what is the relationship between these principles and GDPR?



DIGIPHILE

AI principles not always the same as DP

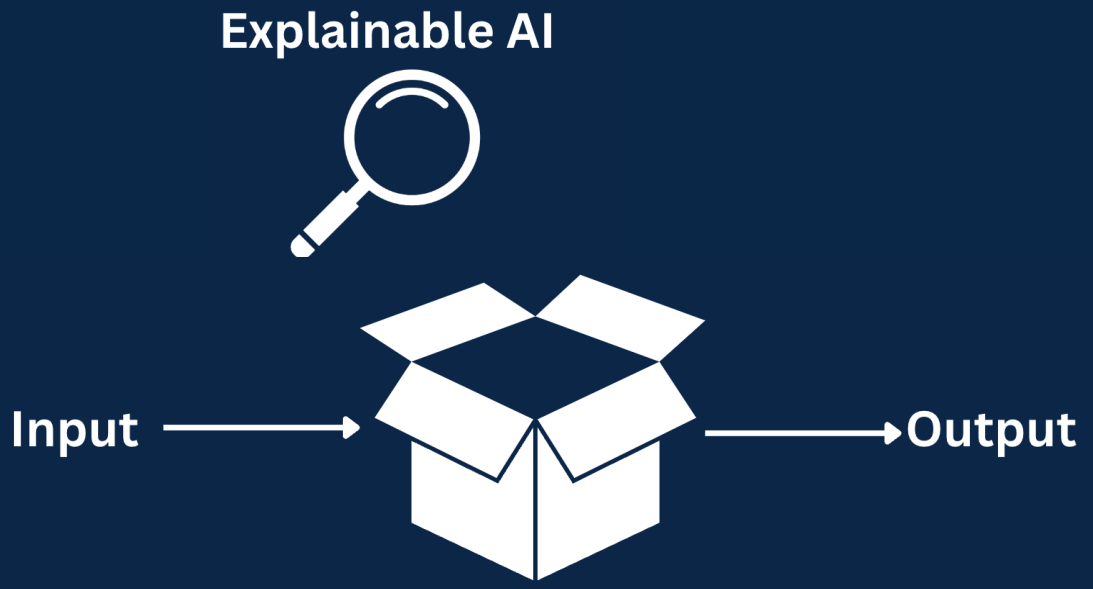
- Task: create an ML-model to predict patient survival rates
- Triage pneumonia patients – keep overnight (inpatient) or send home (outpatient)
- Competition among different ML models. But some odd rules noticed…

Learn more: https://www.pulmonologyadvisor.com/home/topics/practice-management/the-potential-pitfalls-of-machine-learning-algorithms-in-medicine/

DIGIPHILE

Transparency and explainability

- Lawful, fair and transparent processing
- Privacy notice requirements (Art 13 and 14)
- Disclose "*meaningful logic*" behind automated decisions
- From black box to glass box?



DIGIPHILE

Re-purposing data

- Process personal data for "*specified, explicit and legitimate purposes*" …
- Must not be further processed "*in a manner that is incompatible*"
- What is the source of data? What transparency given / rights obtained?
- Processors training ML on controller data – instructions from controller?

DIGIPHILE

Data minimisation v accuracy

- AI needs a lot of data to work
- Less data = less accuracy?
- Tensions with data minimisation
- Possible to use deidentified data?*

* One Gartner predictive report suggests up to 60% of data used to train AI will be synthetic data by 2024: https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601

DIGIPHILE

Bias and discrimination

- Sensitive data = heighted risk:
  - Special category data
  - Criminal convictions data
  - Children and vulnerable people
- Limited grounds to use this data – e.g. Arts 8 – 10 GDPR. But *without* this data, risk bias in AI systems.
- Is there explicit consent? Was data "*manifestly made public by data subject*"? Is data necessary for "*scientific research*"? Other grounds?
- Risk of feedback loop? Could output data serve to bias further training?

DIGIPHILE

Bias and discrimination

"To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data referred to in Article 9(1) of Regulation (EU) 2016/679, Article 10 of Directive (EU) 2016/680 and Article 10(1) of Regulation (EU) 2018/1725, subject to appropriate safeguards..."

Art 10(5) AI Act (Council version)

DIGIPHILE

Bias in AI – an example

China + river =
Yangtze

Paris – France + Italy =
Rome

Doctor – Man + Woman =
Nurse

Programmer – Man + Woman =
Homemaker

From words to phrases and beyond

In certain applications, it is useful to have vector representation of larger pieces of text. For example, it is desirable to have only one vector for representing **'san francisco'**. This can be achieved by pre-processing the training data set to form the phrases using the **word2phrase** tool, as is shown in the example script **./demo-phrases.sh**. The example output with the closest tokens to **'san_francisco'** looks like:

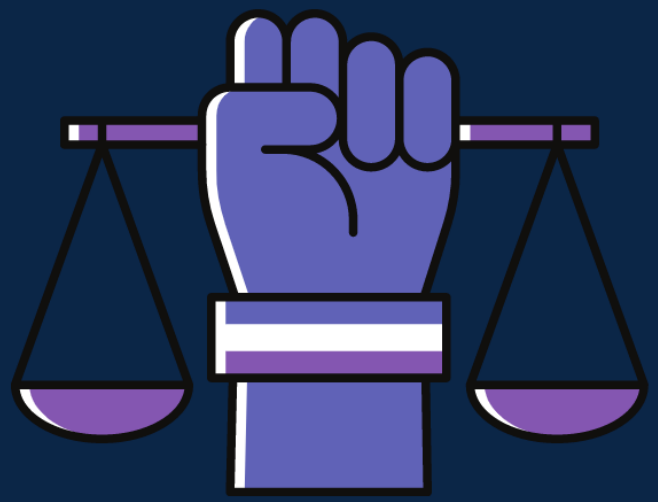
```
Word Cosine distance
------------------------------------------------------------------------
los_angeles                    0.666175
golden_gate                    0.571522
oakland                        0.557521
california                     0.554623
san_diego                      0.534939
pasadena                       0.519115
seattle                        0.512098
taiko                          0.507570
houston                        0.499762
chicago_illinois               0.491598
```

Learn more:
https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

DIGIPHILE

Data subject rights

- GDPR rights to access, correct, delete etc.
- Exercisable against *data*, but what the *model?*
- Right against automated decisions if *legal* effects or *similarly significant* effects...
- ...generally requires human intervention (and, if special category data, explicit consent)



DIGIPHILE

Accountability and governance?

- Who “owns” AI governance?
- Is there a policy framework?
- Is there a DPO?
- Records of processing?
- Privacy by Design and by Default?
- DPIAs?

DIGIPHILE

If you were wondering about that privacy rock song...

(**Verse 1**) I was surfin' the web late one night,
With my heart achin' for some digital delight,
But I couldn't shake this eerie feelin',
That someone was watchin', my privacy stealin'.

They're lurkin' in the shadows, stealin' my soul,
In the darkness, they've taken control,
But I won't let 'em break me, no, not today,
I'm gonna stand up, I'm gonna make them pay.

(**Chorus**) And I'd do anything for privacy (I would do anything),
I'd build a firewall to set me free,
And I'd do anything for privacy (Oh, you know it's true),
But I won't let 'em, no, I won't let 'em take my world from me.

(Verse 2) My identity, it's my own to keep,
But the hackers, they don't ever sleep,
They're tryin' to break down my digital door,
But I'm fightin' back, I won't take it no more.

They say they want to help, but it's just a disguise,
A thousand watchful eyes in the digital skies,
But I won't let them in, I won't let them win,
Together we'll fight, we'll never give in.

(**Chorus**) And I'd do anything for privacy (I would do anything),
I'd build a firewall to set me free,
And I'd do anything for privacy (Oh, you know it's true),
But I won't let 'em, no, I won't let 'em take my world from me.

(**Bridge**) It's a digital battle, a war we're in,
But I'm not alone, I got you by my side,
Together we'll stand, we'll never hide.

(**Verse 3**) The night is long, but the fire burns bright,
We'll stand together, defend our right,
To live our lives without fear or shame,
We won't be just another pawn in their game.

So let's raise our voices, and let 'em know,
We won't be silenced, we won't let go,
The fight for privacy, it starts today,
We'll stand together, we'll make 'em pay.

(Chorus) And I'd do anything for privacy (I would do anything),
I'd build a firewall to set me free,
And I'd do anything for privacy (Oh, you know it's true),
But I won't let 'em, no, I won't let 'em take my world from me.

DIGIPHILE

(A selection of) further reading for EU/UK professionals

- ICO: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/
- CNIL: https://www.cnil.fr/en/artificial-intelligence-cnil-publishes-set-resources-professionals
- AEPD / EDPS: https://www.aepd.es/es/documento/10-misunderstandings-machinelearning-en.pdf
- EDPB: https://edpb.europa.eu/our-work-tools/our-documents/topic/artificial-intelligence_en

Phil Lee (Managing Director)

Digiphile - Simple. Strategic. Actionable.

+44 (0) 7598 245865

phil.lee@digiphile.law

www.digiphile.law

* No robots were harmed in the production of these slides.