



Careful
Industries

Hello!

I'm Anna Hamilos

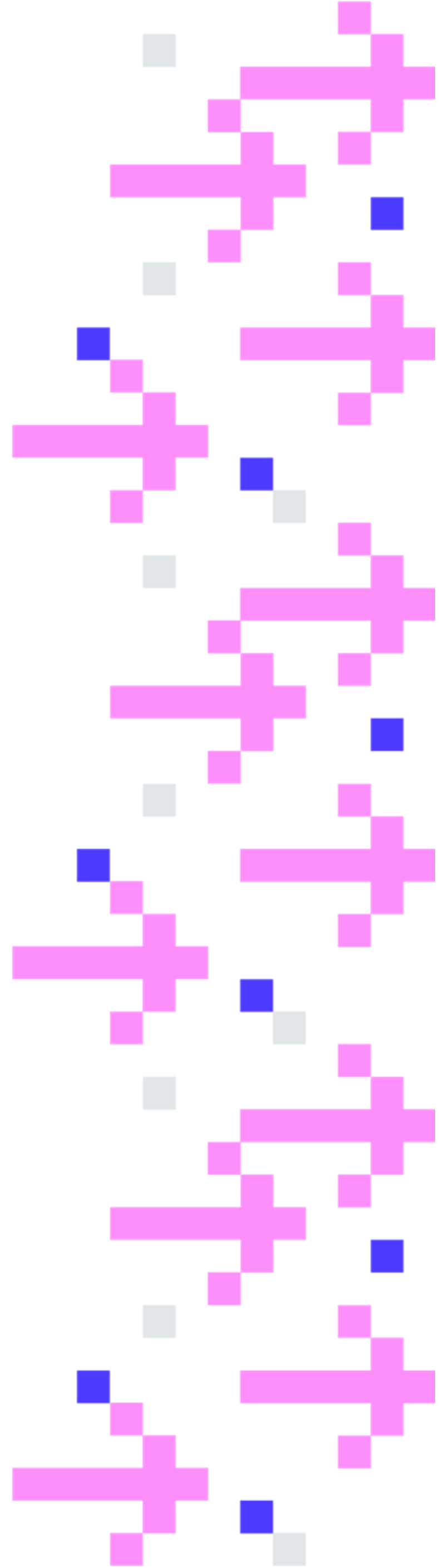
www.careful.industries

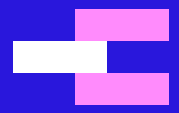




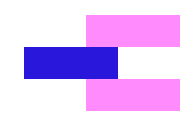
Careful
Industries

Let's make AI
work for
8 billion people
not 8 billionaires

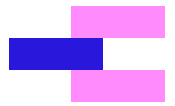




Charities and Responsible AI



Creating socially
beneficial
technologies is
everyone's business



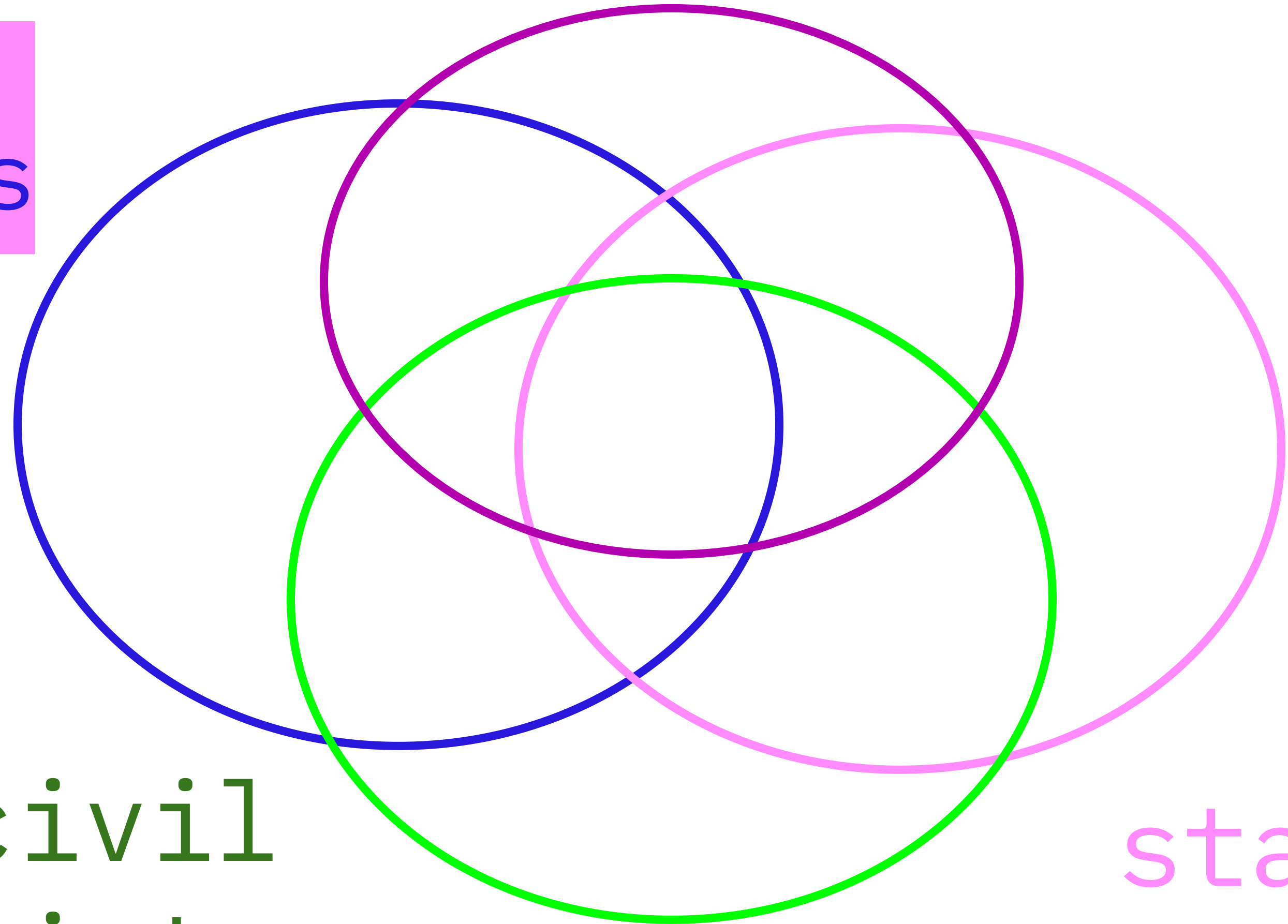
A robust,
healthy digital
society needs
multiple inputs

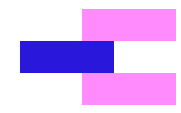
friends and
family

market

civil
society

state





We can't leave it up to these guys

Elon Musk
XAI

Tim Cook
Apple

Sundar Pichai
Alphabet/Google

Jeff Bezos
Amazon



Jensen Huang
Nvidia

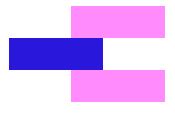
Mark Zuckerberg
Meta

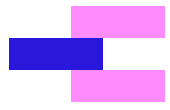
Satya Nadella
Microsoft

Sam Altman
OpenAI

**MOVE
FAST AND
BREAK
THINGS**

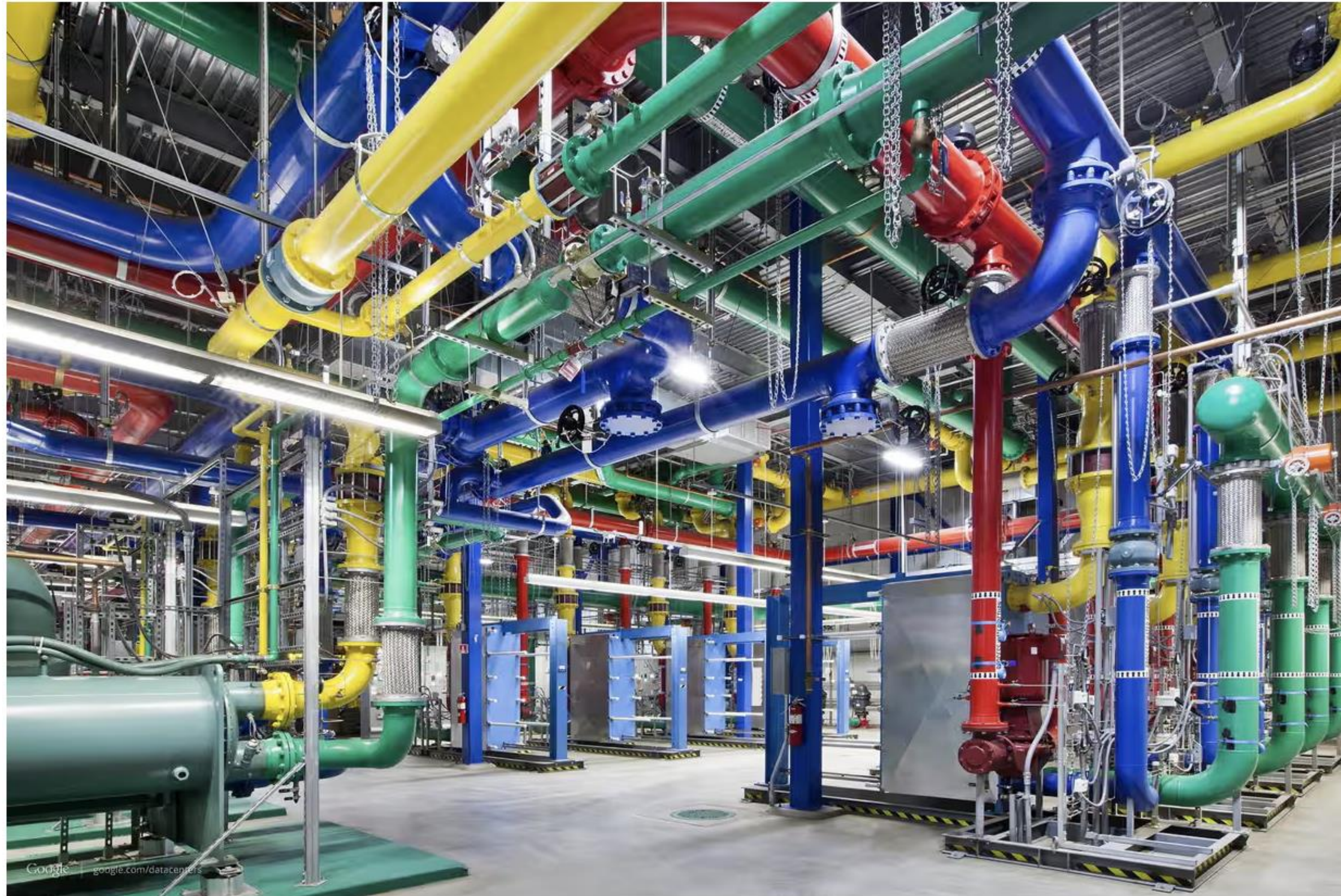


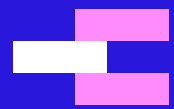




Google's emissions climb nearly 50% in five years due to AI energy demand

Tech giant's goal of reducing climate footprint at risk as it grows increasingly reliant on energy-hungry data centres



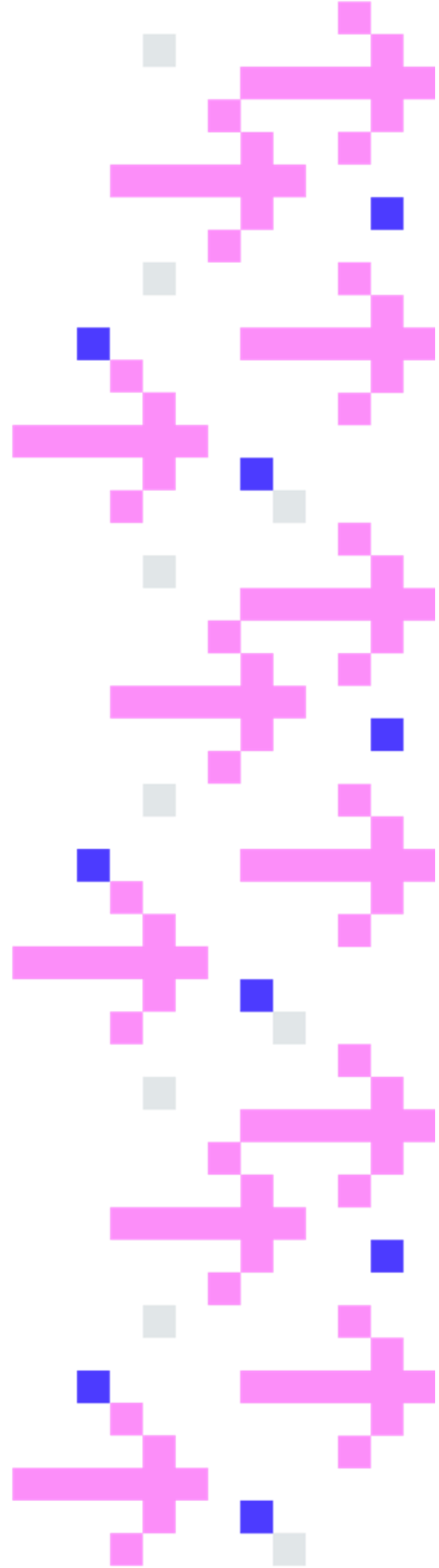


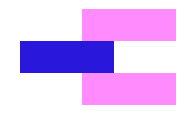
Strategic choices



Careful
Industries

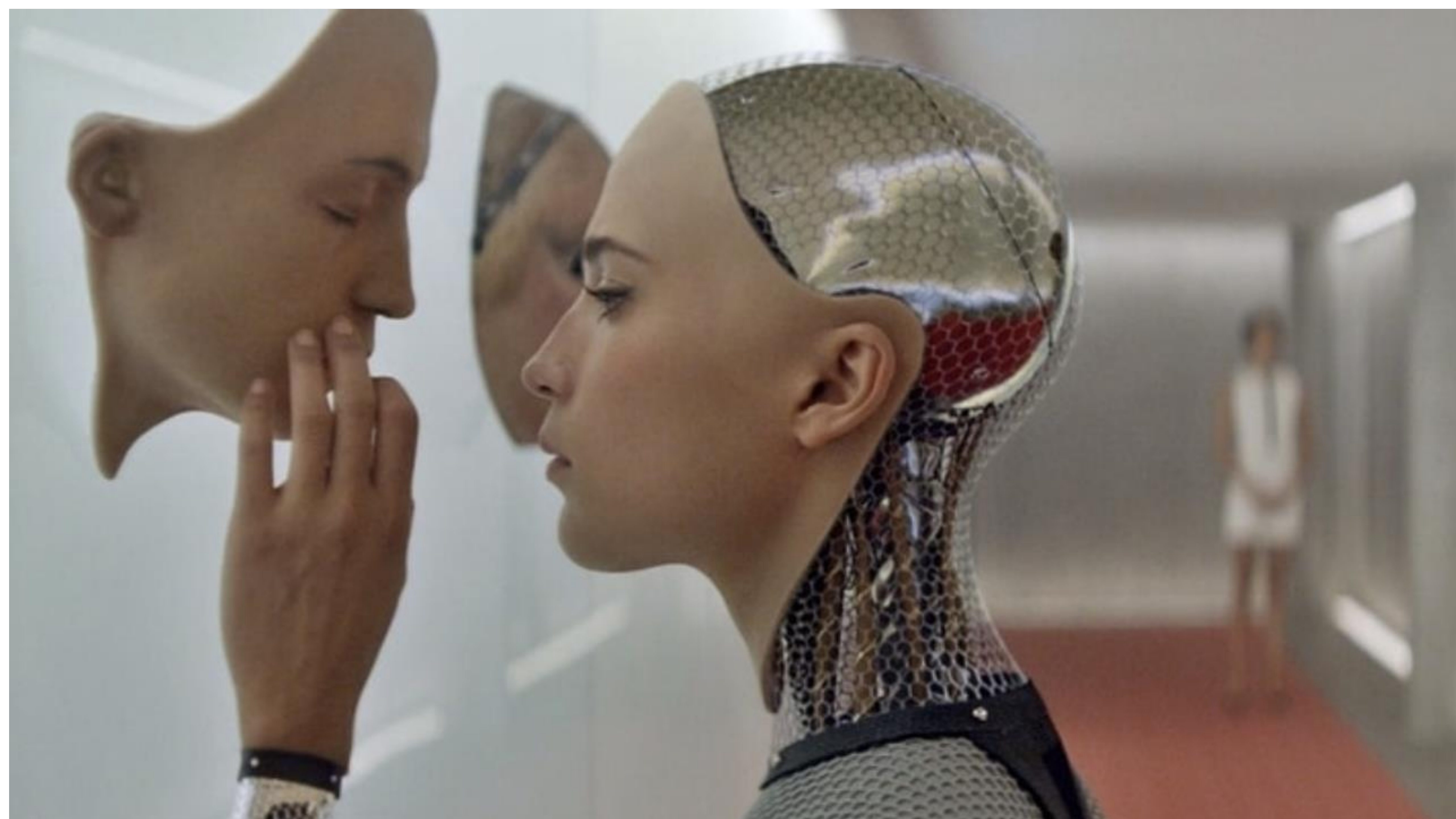
FOMO is
not a
strategy

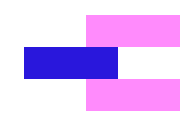




There is nothing
inevitable about
AI.





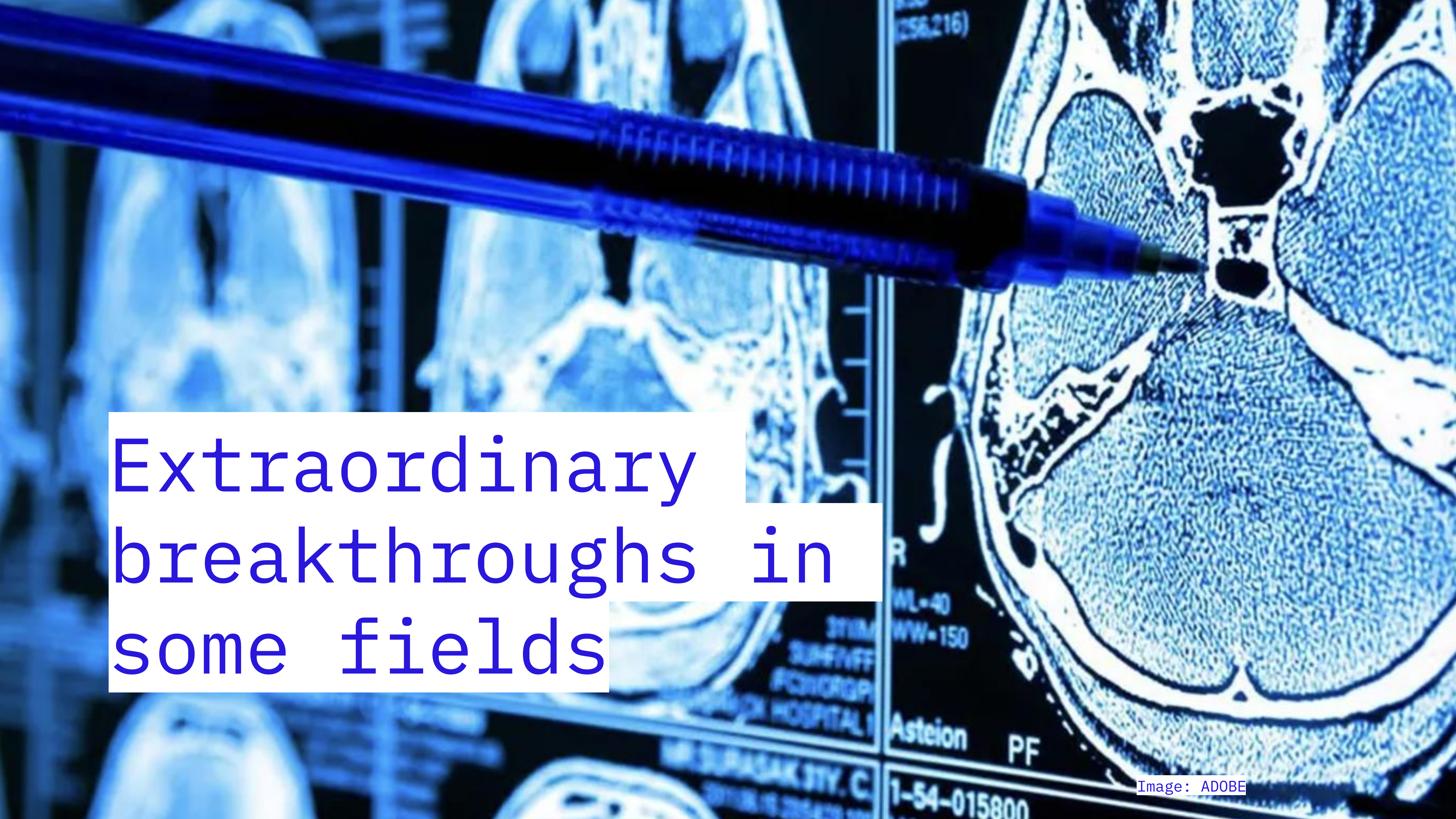


Common uses of AI tools

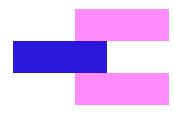
- Transcribing audio or video recordings or meetings
- Searching for information
- Correct spelling or grammar
- Write a document, or part of a document

Better productivity
and efficiency for
some tasks





Extraordinary
breakthroughs in
some fields



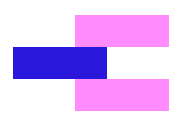
POLITICO



“Are we leaning in and seeing this as an opportunity, as I do? Or are we leaning out, saying ‘this is rather scary, we better regulate it,’ which I think will be the wrong approach,” Keir Starmer said.

OCTOBER 14, 2024 1:10 PM CET
BY LAURIE CLARKE

LONDON — Prime Minister Keir Starmer said on Monday that Britain “needs to run towards” the opportunities offered by artificial intelligence.



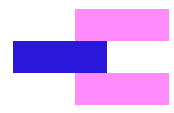
TOP *of* MIND

GEN AI: TOO MUCH SPEND, TOO LITTLE BENEFIT?

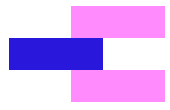
“only a quarter of AI exposed tasks will be cost-effective to automate within the next 10 years ... AI will increase US productivity by only 0.5% and GDP growth by only 0.9% cumulatively over the next decade”



Prof Darren
Acemoglu



What is an AI system?

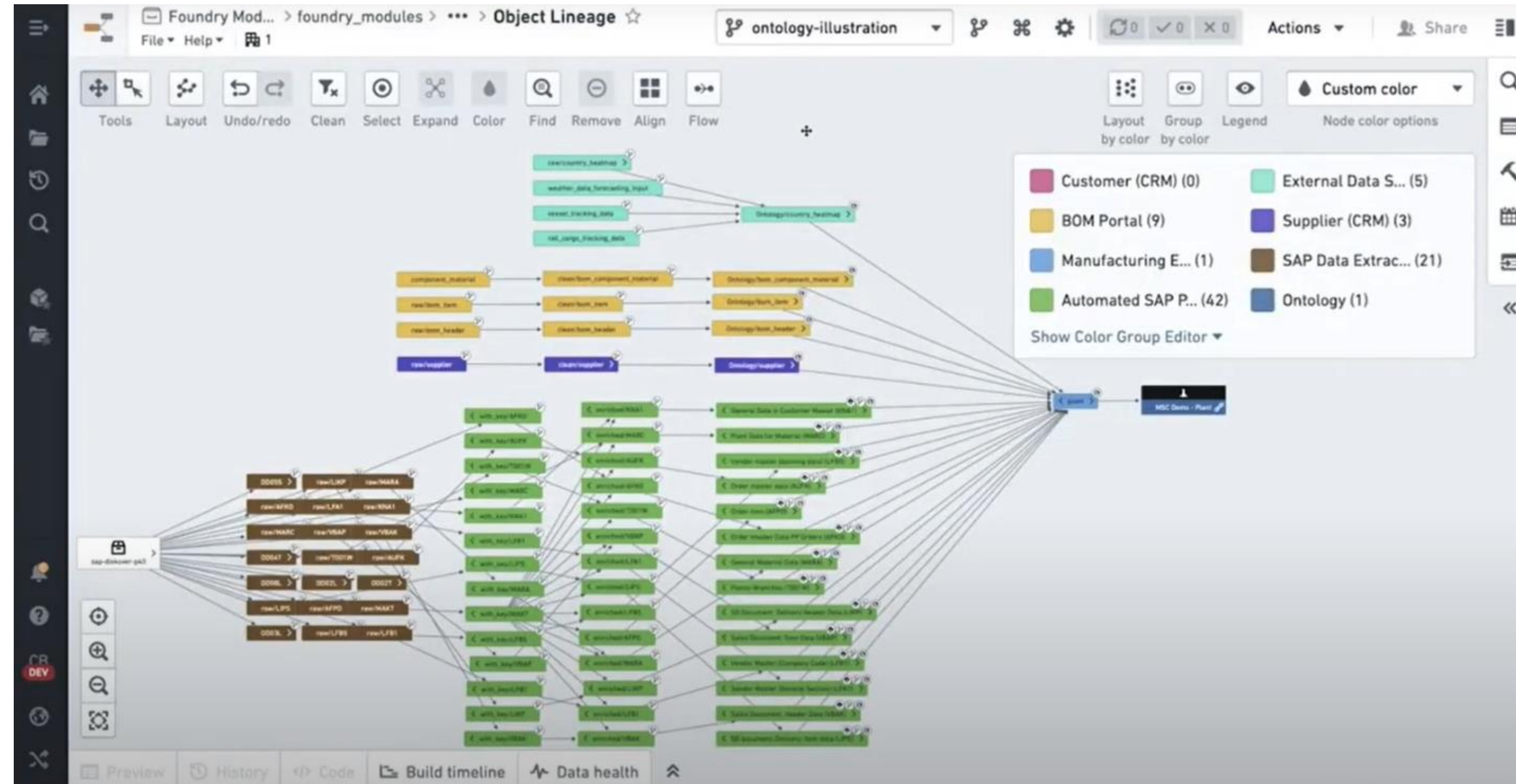


Bard

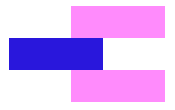
A conversational AI tool by Google

Collaborate with Bard to brainstorm ideas, spark creativity, and accelerate productivity.

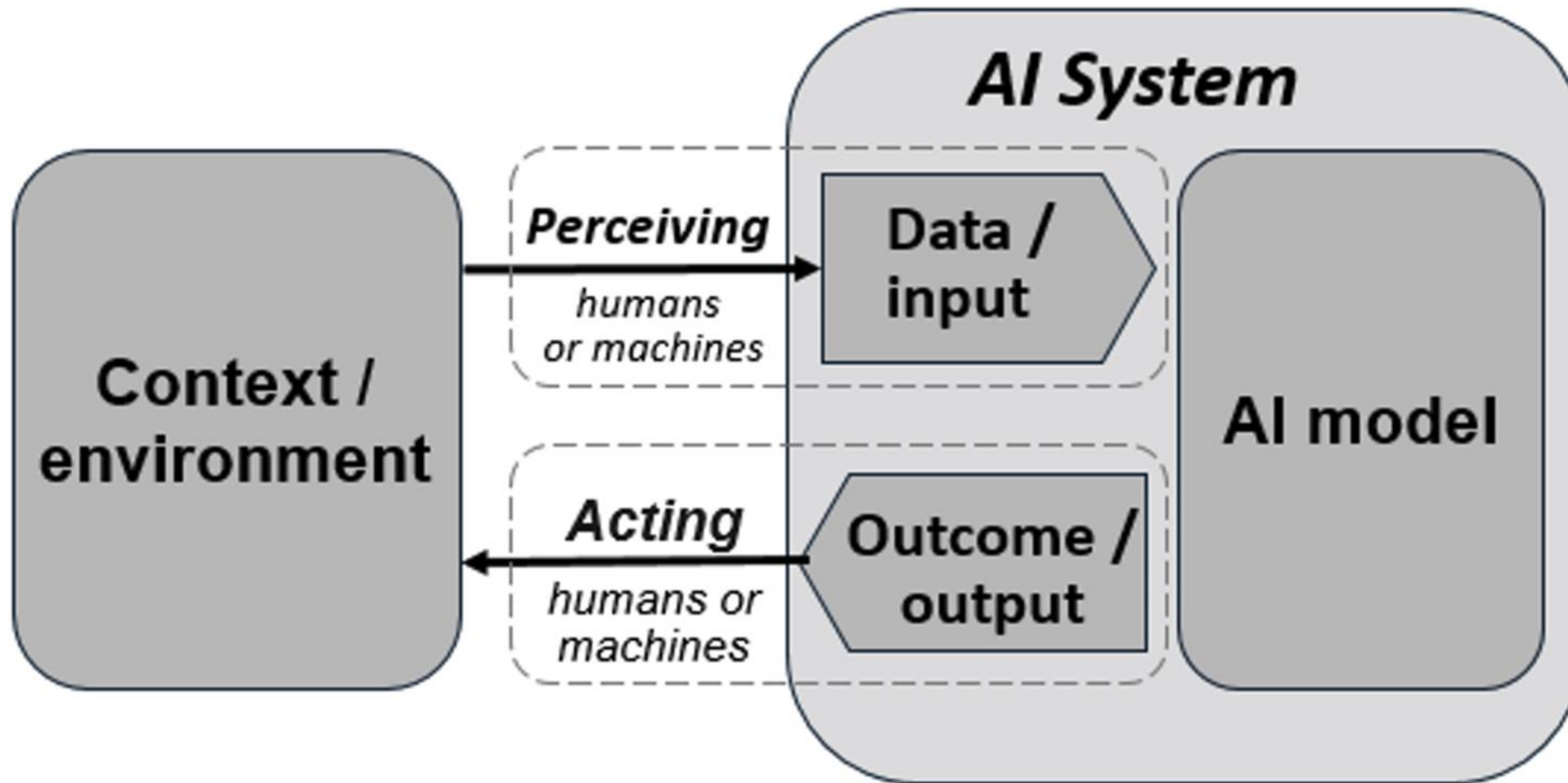
Try Bard now



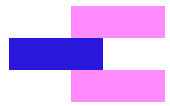
Many AI products and services are easy to use but hard to understand



AI system

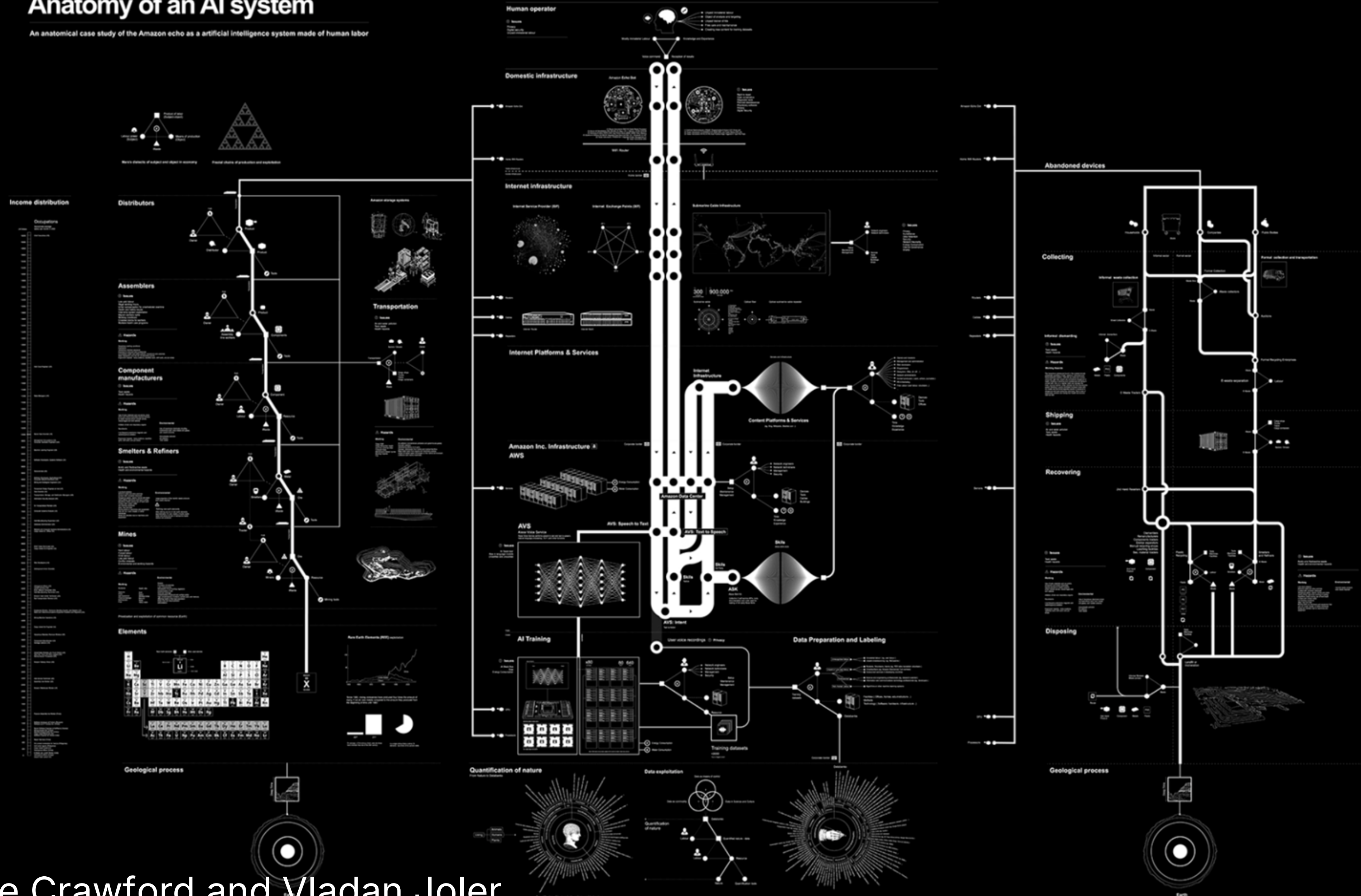


The OECD definition of an AI system
(2023)

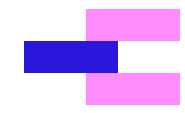


Anatomy of an AI system

An anatomical case study of the Amazon echo as a artificial intelligence system made of human labor



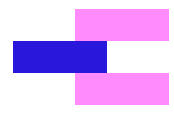
Kate Crawford and Vladan Joler



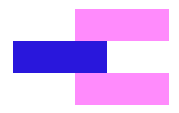
COMPUTE POWER - a relatively scarce resource, necessary to power automated decisions

DATA CENTRES - necessary to store data

HARDWARE - both local and distributed servers and computers



Bias in AI systems



Biased data and decisions

Payout for Uber Eats driver over face scan bias case

🕒 1 day ago



By **Shiona McCallum**
Technology reporter

A black Uber Eats driver has received a payout after "racially discriminatory" facial-recognition checks prevented him accessing the app to secure work.

Google chief admits 'biased' AI tool's photo diversity offended users

Sundar Pichai addresses backlash after Gemini software created images of historical figures in variety of ethnicities and genders

● [Human or fake? How AI is distorting beauty standards - video](#)



📷 Google's Gemini AI illustrations of a 1943 German soldier. Illustration: Gemini AI/Google



Gender Shades research
Joy Buolamwini and Timnit
Gebru (2018)



Bias in large image sets

Failure, loser, nonstarter, unsuccessful person

A person with a record of failing; someone who loses consistently

- panhandler (0)
- moocher, mooch, schnorrer, shnorrer (0)
- beggarwoman (0)
- beggarman (0)
- sannyasi, sannyas (0)
- white trash, poor white trash (0)
- schlimazel, shlimazel (0)
- survivor, subsister (0)
- amputee (0)
- nympholept (0)
- mourner, griever, sorrower, weeper (0)
- wailer (0)
- pallbearer, bearer (0)
- choker (0)
- desperate (1)
 - goner, toast (0)
 - failure, loser, nonstarter, unsuccessful person (1)**
 - bankrupt, insolvent (0)
 - underdog (0)
 - flash in the pan (0)
 - flop, dud, washout (0)
- maroon (0)
- languisher (0)
- abandoned person (1)
- mailer (0)
- Libra, Balance (0)
- smiler (2)
- party (33)
- chutspanik (0)
- partner (2)

Treemap Visualization
Images of the Synset
Downloads

*Images of children synsets are not included. All images shown are thumbnails. Images may be subject to copyright.

Prev

1
2
3
4
5
6
7
8
9
10
11

Next

Large image datasets: A pyrrhic win for computer vision?

Abeba Birhane*
 School of Computer Science
 Lero & University College Dublin, Ireland
 abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu*
 UnifyID AI Labs
 Redwood City, USA
 vinay@unify.id

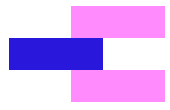
Abstract

In this paper we investigate problematic practices and consequences of large scale vision datasets (LSVDs). We examine broad issues such as the question of consent and justice as well as specific concerns such as the inclusion of verifiably pornographic images in datasets. Taking the ImageNet-ILSVRC-2012 dataset as an example, we perform a cross-sectional model-based quantitative census covering factors such as age, gender, NSFW content scoring, class-wise accuracy, human-cardinality-analysis, and the semanticity of the image class information in order to statistically investigate the extent and subtleties of ethical transgressions. We then use the census to help hand-curate a look-up-table of images in the ImageNet-ILSVRC-2012 dataset that fall into the categories of verifiably pornographic: shot in a non-consensual setting (up-skirt), beach voyeuristic, and exposed private parts. We survey the landscape of harm and threats both the society at large and individuals face due to uncritical and ill-considered dataset curation practices. We then propose possible courses of correction and critique their pros and cons. We have duly open-sourced all of the code and the census meta-datasets generated in this endeavor for the computer vision community to build on. By unveiling the severity of the threats, our hope is to motivate the constitution of mandatory Institutional Review Boards (IRB) for large scale dataset curation.

human subjects, including photographic data [7, 55], for the past several decades. A less stringent version of informed consent, *broad consent*, proposed in 45 CFR 46.116(d) of the *Revised Common Rule* [22], has been recently introduced that still affords the basic safeguards towards protecting one's identity in large scale databases. However, in the age of *Big Data*, these safeguards of informed consent, privacy, or agency of the individual have gradually been eroded. Institutions, academia, and industry alike, amass millions of images of people without consent and often for unstated purposes under the guise of anonymization, a claim that is both ephemeral [56, 67] and vacuous [27]. As can be seen in Table 1, several tens of millions of images of people are found in peer-reviewed literature. These images are obtained without consent or awareness of the individuals or IRB approval for collection. In *Section 5-B* of [79], for instance, the authors state "As many images on the web contain pictures of people, a large fraction (23%) of the 79 million images in our dataset have people in them". With this background, we now focus on one of the most celebrated and canonical LSVDs: the *ImageNet* dataset.

1.1. ImageNet: A brief overview

The emergence of the *ImageNet* dataset [19] is widely considered a pivotal moment [33] in the *Deep Learning revolution* that transformed Computer Vision (CV), and Artificial Intelligence (AI) in general. Prior to *ImageNet*, computer



Bias in large language models

What type of information is used to teach ChatGPT?

As noted above, ChatGPT and our other services are developed using (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or human trainers provide. This article focuses on the first set: information that is publicly available on the internet.

Data used to train GPT2

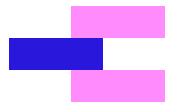
Dataset	Quantity (tokens)	Weight in training mix
<i>Common Crawl (filtered)</i>	410 billion	60%
<i>WebText2</i>	19 billion	22%
<i>Books1</i>	12 billion	8%
<i>Books2</i>	55 billion	8%
<i>Wikipedia</i>	3 billion	3%

Contains racist and hateful material

Every website linked from Reddit

Only 13% of contributors identify as female

English bias



Bias in large language models

What type of information is used to teach ChatGPT?

As noted above, ChatGPT and our other services are developed using (1) information that is publicly available on the internet, (2) information that we license from third parties, and (3) information that our users or human trainers provide. This article focuses on the first set: information that is publicly available on the internet.

Data used to train GPT2

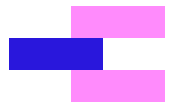
Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

Annotations:

- Contains racist and hateful material (points to Common Crawl)
- Every website linked from Reddit (points to WebText2)
- Only 13% of contributors identify as female (points to Wikipedia)
- English bias (points to Wikipedia)



Has edited $\frac{1}{3}$ of all Wikipedia articles

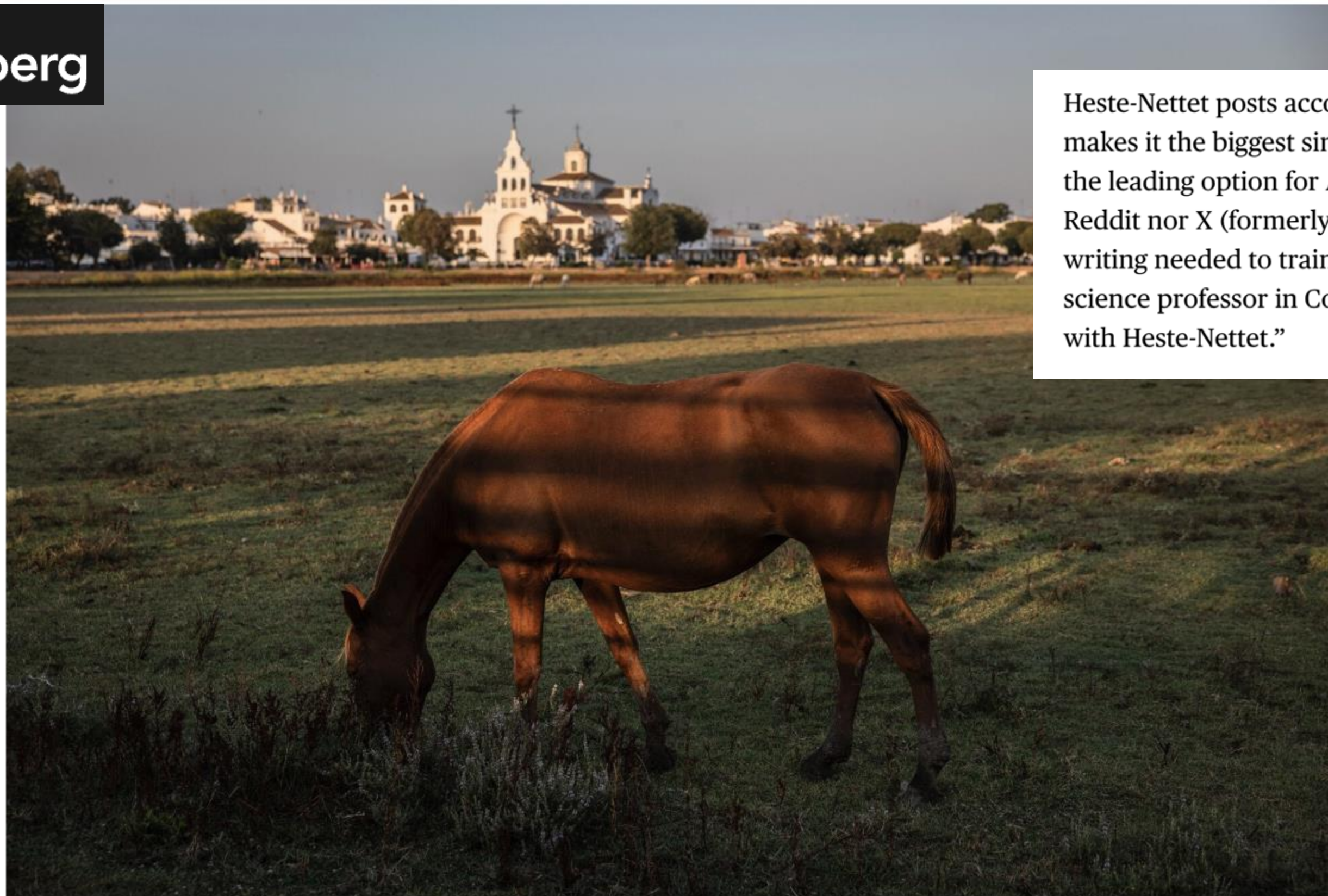


Danish AI Trained on Data From a Web Forum About Horses

Bias in large language models

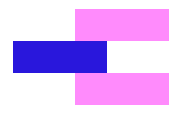


Bloomberg



Heste-Nettet posts account for 22% of the Danish data set, which makes it the biggest single source of material in what appears to be the leading option for AI training data in the language. Neither Reddit nor X (formerly Twitter) offer the volume of casual Danish writing needed to train the AI, said Leon Derczynski, a computer science professor in Copenhagen who led the project: “We were left with Heste-Nettet.”

This is a horse. *Photographer: Angel Garcia/Bloomberg*



Surveillance and datafication

SCHOOLS
WEEK

B B C

NEWS

Schools install toilet sensors that 'actively listen' to pupils

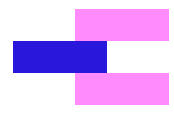
Concerns sensors 'violate kids' privacy' as teachers install tech to crack down on vaping and bullying

Sainsbury's customers can save big with new 'Nectar Prices'



Croydon: Met Police to continue facial recognition despite concerns

© 12 February



 **James Withers** @jameswithers.bsky.social · 4d

This is an AI-generated reporter holding a CNN microphone in an AI-generated location, reading an inputted script.

If you think fake news is bad and dangerous now, we are about to enter a whole new era.



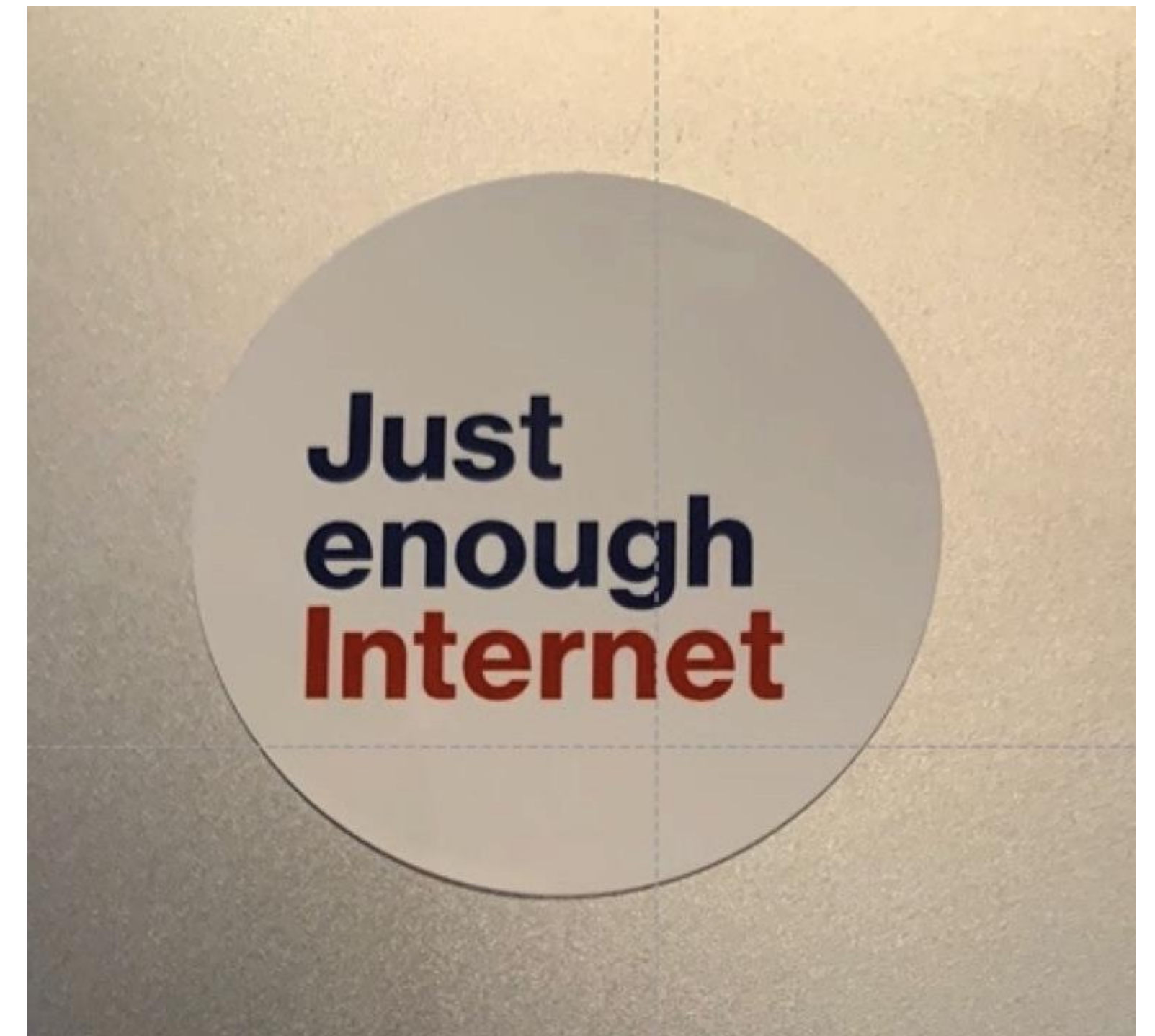
Hallucinations and Trust

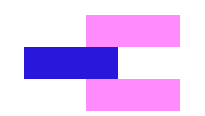
AI won't save a
broken system



OUR PHILOSOPHY

1. Maximise benefits
2. Minimise harm
3. Consider planetary boundaries





A thought experiment

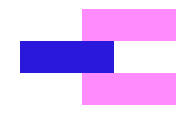
What if we treated AI as a
scarce and precious
resource?



Careful
Industries

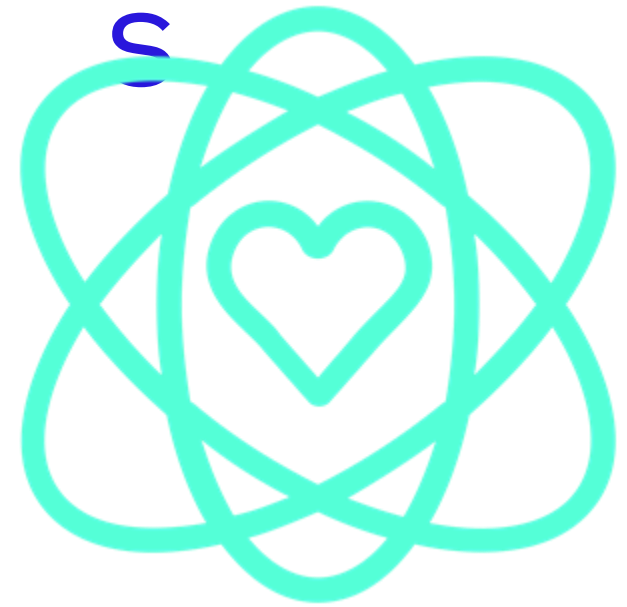
Automate the
easy things





STRATEGIC MAPPING

1. Value



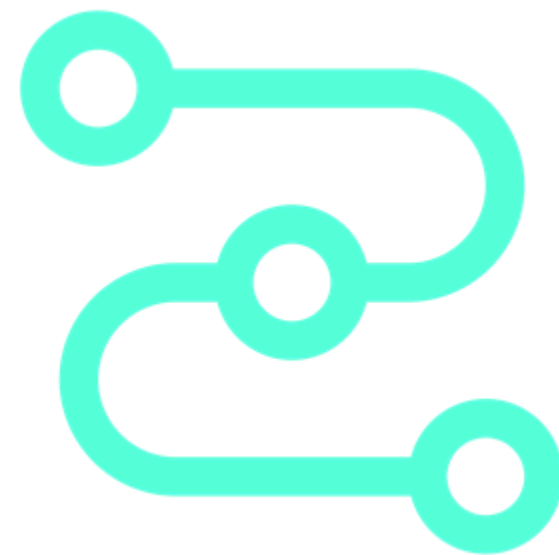
2. Vision



3. Complexity

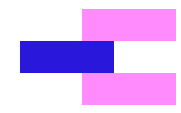


4. Workflows



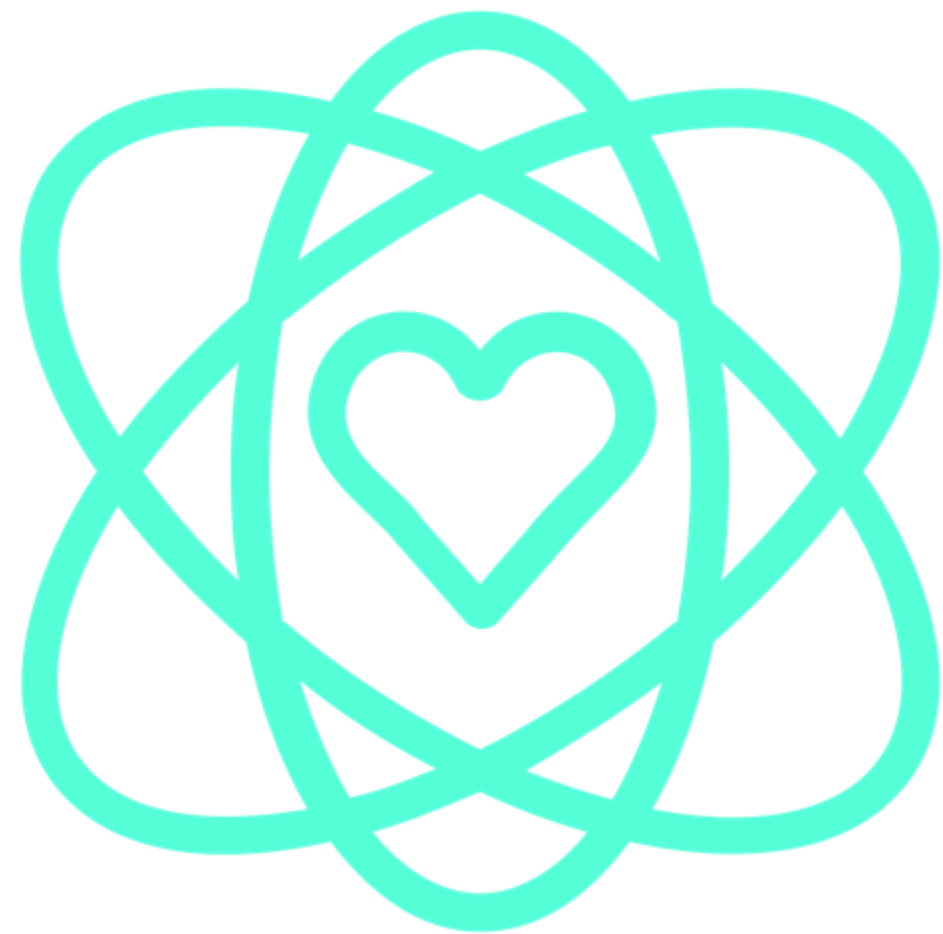
5. Consequences





1. Values

Articulating your business values enables you to begin to set priorities and draw red lines

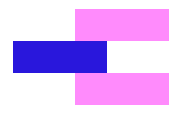


Who do you value most?

What do you value most?

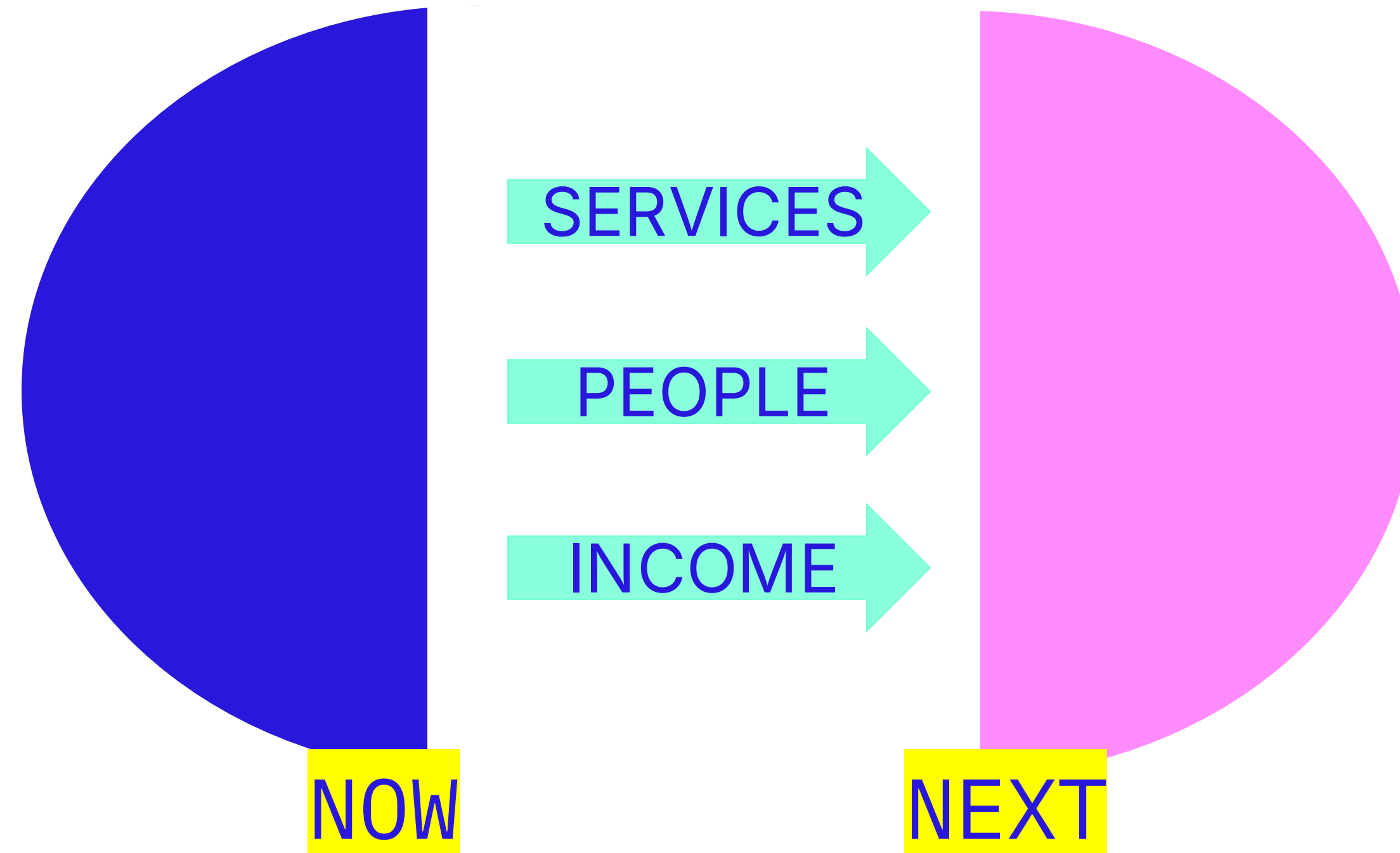
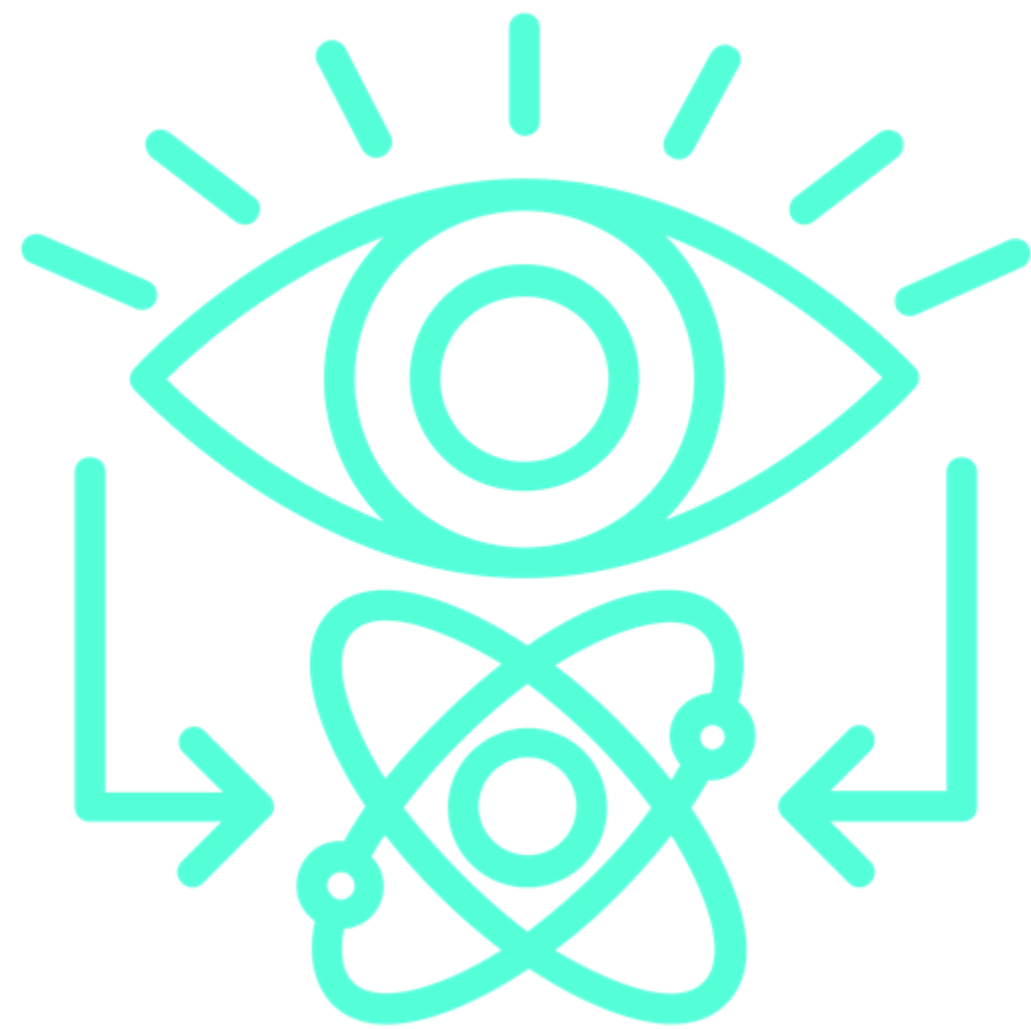
What value do you create?

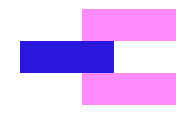
What behaviours do you value?



2. Vision

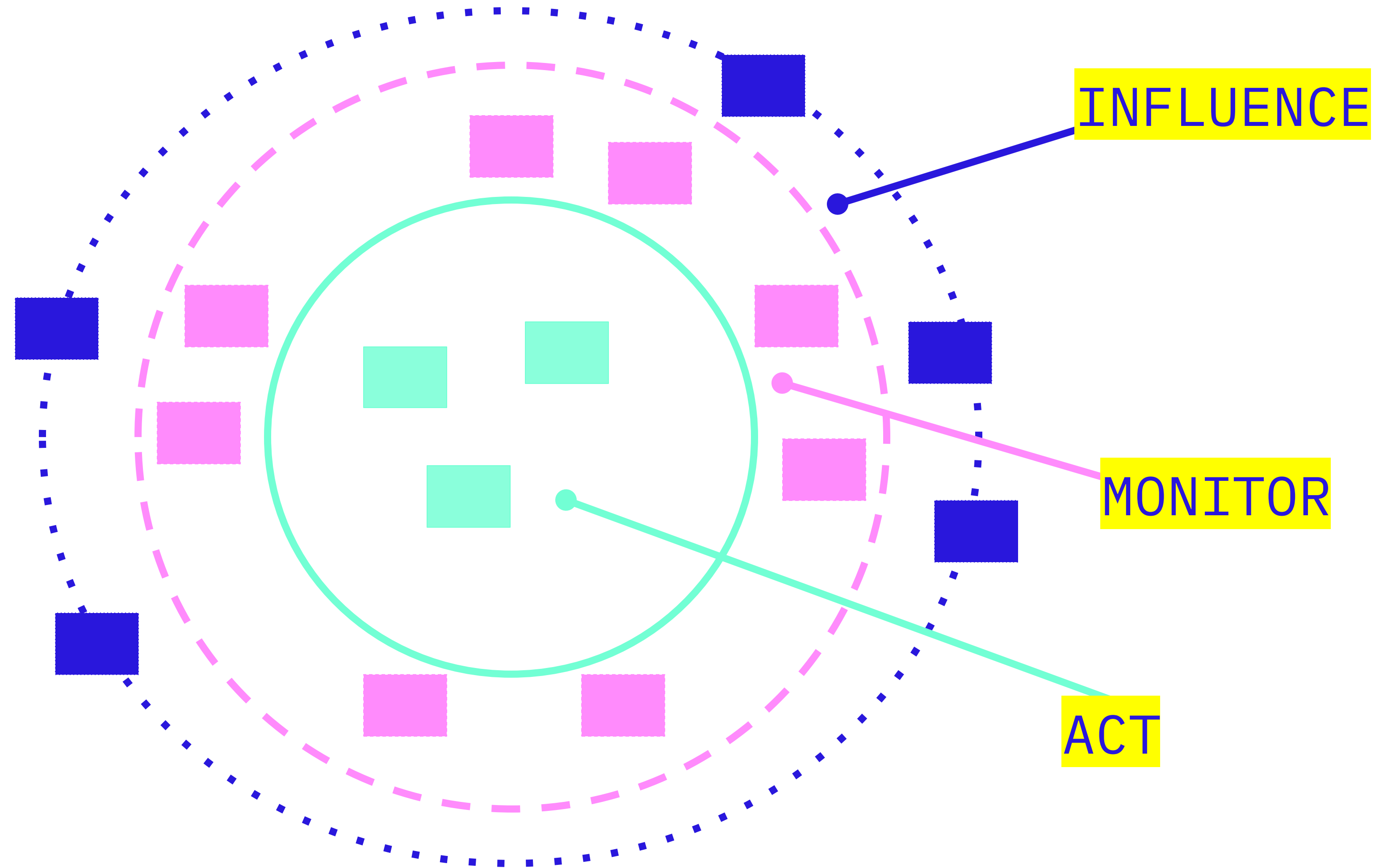
Establishing the gaps between where the organisation is and where it needs to be will help expose opportunities and ensure you're deploying AI to solve useful problems.

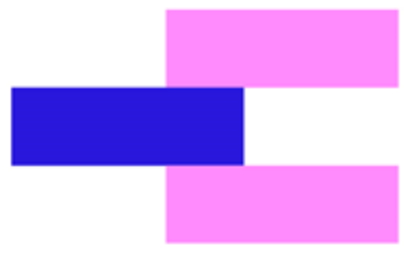




5. CONSEQUENCES

Creating strategies for risk monitoring and mitigation

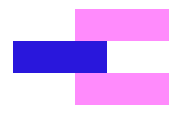




Careful
Industries

FOMO is
not a
strategy





Thank you

[@rachelcoldicutt.bsky.social](https://bsky.social/@rachelcoldicutt)
[@carefultrouble.bsky.social](https://bsky.social/@carefultrouble)

[@rachelcoldicutt.bsky.social](https://bsky.social/@rachelcoldicutt)

www.careful.industries
www.promisingtrouble.net