

Environmental, Ethical, and Societal Implications of AI

Steven Schramm
University of Geneva

2 June 2026

IN2P3 School of Statistics

(Image by ChatGPT)



Introduction

Artificial Intelligence (AI) is an incredibly powerful tool

- Already many stories on how those who don't use AI will be left behind
- Also stories of how AI is going to be the "next industrial revolution"

Introduction

Artificial Intelligence (AI) is an incredibly powerful tool

- Already many stories on how those who don't use AI will be left behind
- Also stories of how AI is going to be the "next industrial revolution"

AI certainly has transformative potential, also known as a "disruptive technology"

- An innovation that fundamentally alters established/legacy systems
- Implications across the spectrum: high-energy physics to everyday life and more
- We are living in an age where AI may fundamentally alter the future of humanity

It is easy enough to give a talk on the opportunities that AI presents

- Tune into any new AI model announcement and you'll get that

Introduction

Artificial Intelligence (AI) is an incredibly powerful tool

- Already many stories on how those who don't use AI will be left behind
- Also stories of how AI is going to be the "next industrial revolution"

AI certainly has transformative potential, also known as a "disruptive technology"

- An innovation that fundamentally alters established/legacy systems
- Implications across the spectrum: high-energy physics to everyday life and more
- We are living in an age where AI may fundamentally alter the future of humanity

It is easy enough to give a talk on the opportunities that AI presents

- Tune into any new AI model announcement and you'll get that

However, that sweeps under the rug a lot of important implications of AI usage

- The development and use of AI comes with great costs and risks
- That is the subject of this talk: ecological, ethical, and society implications of AI

Who am I to discuss this topic?

I am a particle physicist, working on collider and gravitational-wave physics
- On what basis am I discussing the impact of AI beyond physics?

Who am I to discuss this topic?

I am a particle physicist, working on collider and gravitational-wave physics
- On what basis am I discussing the impact of AI beyond physics?

I have a history of inter-disciplinary and trans-disciplinary AI discussions
- I am one of the founders of the CERN inter-experiment ML group (2015)
- Also a founder of symposium on AI for industry, science, and society ('19, '21, '23)
- I am on the University of Geneva's Faculty of Science committee on use of Generative AI



Who am I to discuss this topic?

I am a particle physicist, working on collider and gravitational-wave physics

- On what basis am I discussing the impact of AI beyond physics?

I have a history of inter-disciplinary and trans-disciplinary AI discussions

- I am one of the founders of the CERN inter-experiment ML group (2015)
- Also a founder of symposium on AI for industry, science, and society ('19, '21, '23)
- I am on the University of Geneva's Faculty of Science committee on use of Generative AI



This is a topic I consider important, but it is not my job

- I am an advocate for more discourse between AI experts and societal stakeholders
- I am here to give you a flavour of a quickly evolving discussion

With that out of the way, let us dive into the implications of using AI

...but first off, what is AI?

Unfortunately, nomenclature can be very confusing:

- **AI** = **A**rtificial **I**ntelligence
 - Algorithms that can perform tasks typically associated with human intelligence
- **ML** = **M**achine **L**earning
 - A subset of **AI**, learning patterns from data and generalise to unseen data

...but first off, what is AI?

Unfortunately, nomenclature can be very confusing:

- **AI** = **A**rtificial **I**ntelligence
 - Algorithms that can perform tasks typically associated with human intelligence
- **ML** = **M**achine **L**earning
 - A subset of **AI**, learning patterns from data and generalise to unseen data
- **AGI** = **A**rtificial **G**eneral **I**ntelligence
 - A hypothetical **AI** that matches/surpasses humans across a spectrum of tasks

...but first off, what is AI?

Unfortunately, nomenclature can be very confusing:

- **AI** = **A**rtificial **I**ntelligence
 - Algorithms that can perform tasks typically associated with human intelligence
- **ML** = **M**achine **L**earning
 - A subset of **AI**, learning patterns from data and generalise to unseen data
- **AGI** = **A**rtificial **G**eneral **I**ntelligence
 - A hypothetical **AI** that matches/surpasses humans across a spectrum of tasks

In physics, we primarily use **ML** - advanced pattern recognition

- We have a given task that we want an algorithm to learn to perform

...but first off, what is AI?

Unfortunately, nomenclature can be very confusing:

- **AI** = **A**rtificial **I**ntelligence
 - Algorithms that can perform tasks typically associated with human intelligence
- **ML** = **M**achine **L**earning
 - A subset of **AI**, learning patterns from data and generalise to unseen data
- **AGI** = **A**rtificial **G**eneral **I**ntelligence
 - A hypothetical **AI** that matches/surpasses humans across a spectrum of tasks

In physics, we primarily use **ML** - advanced pattern recognition

- We have a given task that we want an algorithm to learn to perform

When I say **AI**, I typically refer to tasks that do not fall under **ML**

- Many of the following also applies to **ML**, but **AI** increases the importance
- In reality, almost everything referred to as "**AI**" at this point as it's a big buzzword

An example of ML vs more general AI

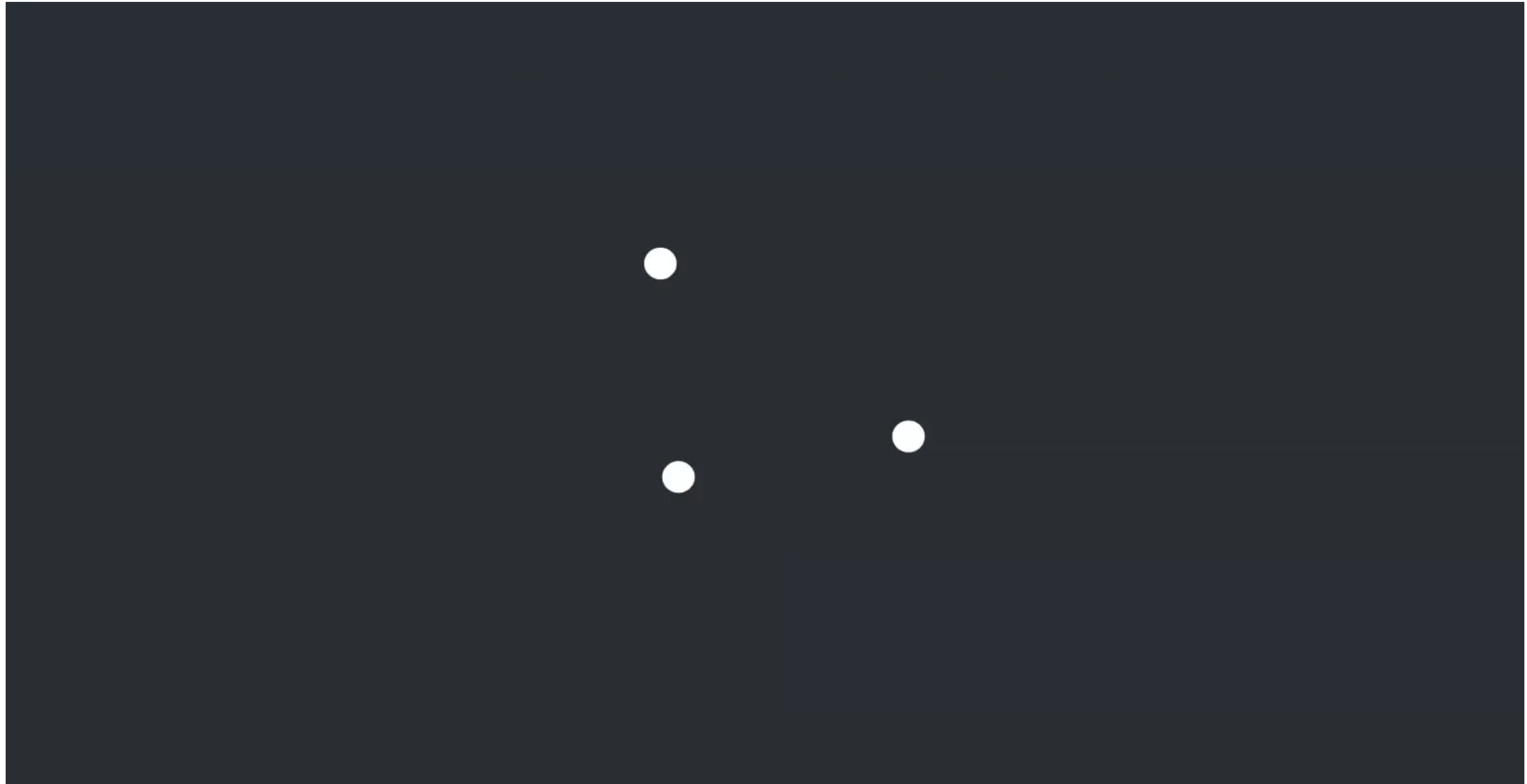
Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)

AI: MIMIC THE HUMAN BRAIN



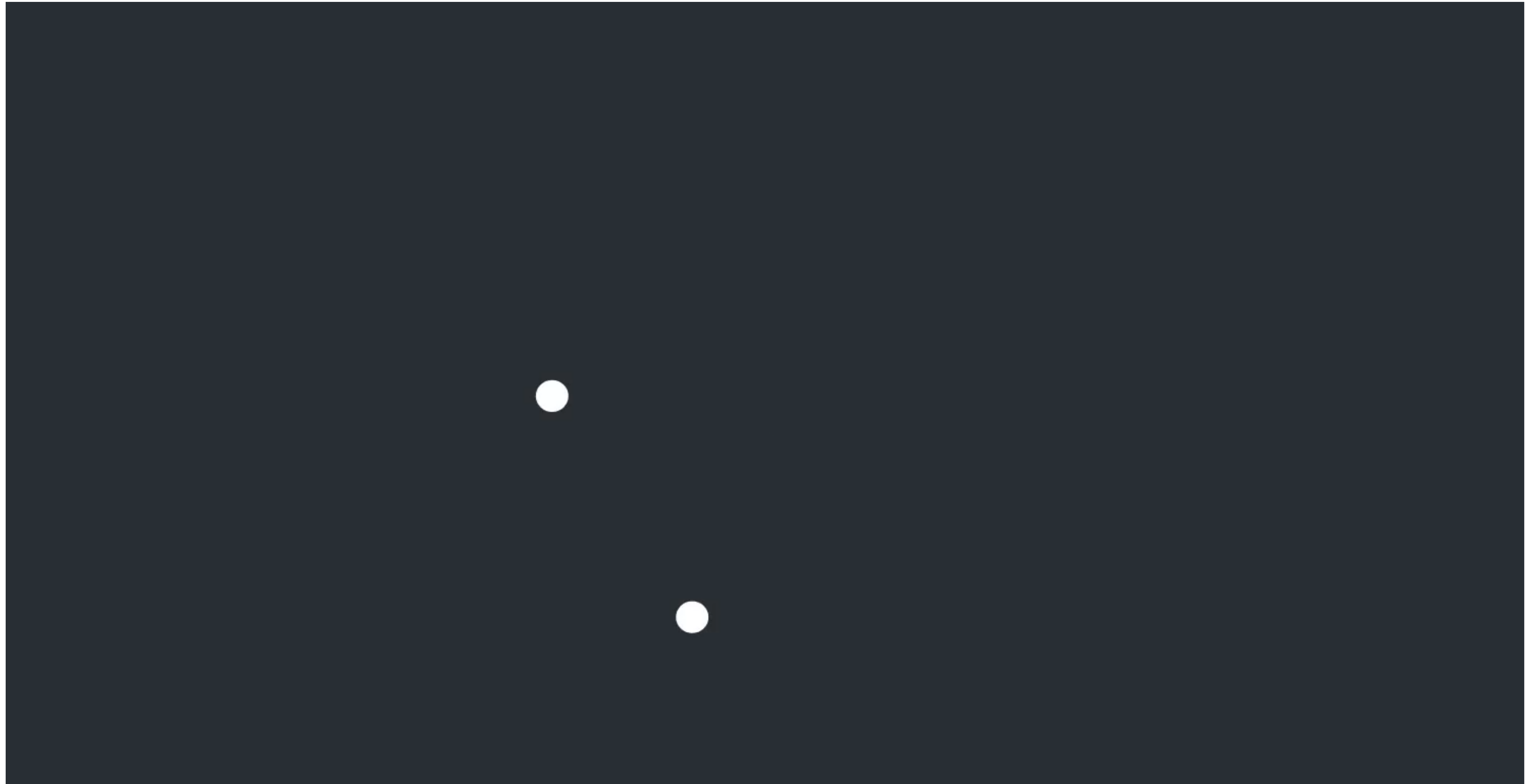
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



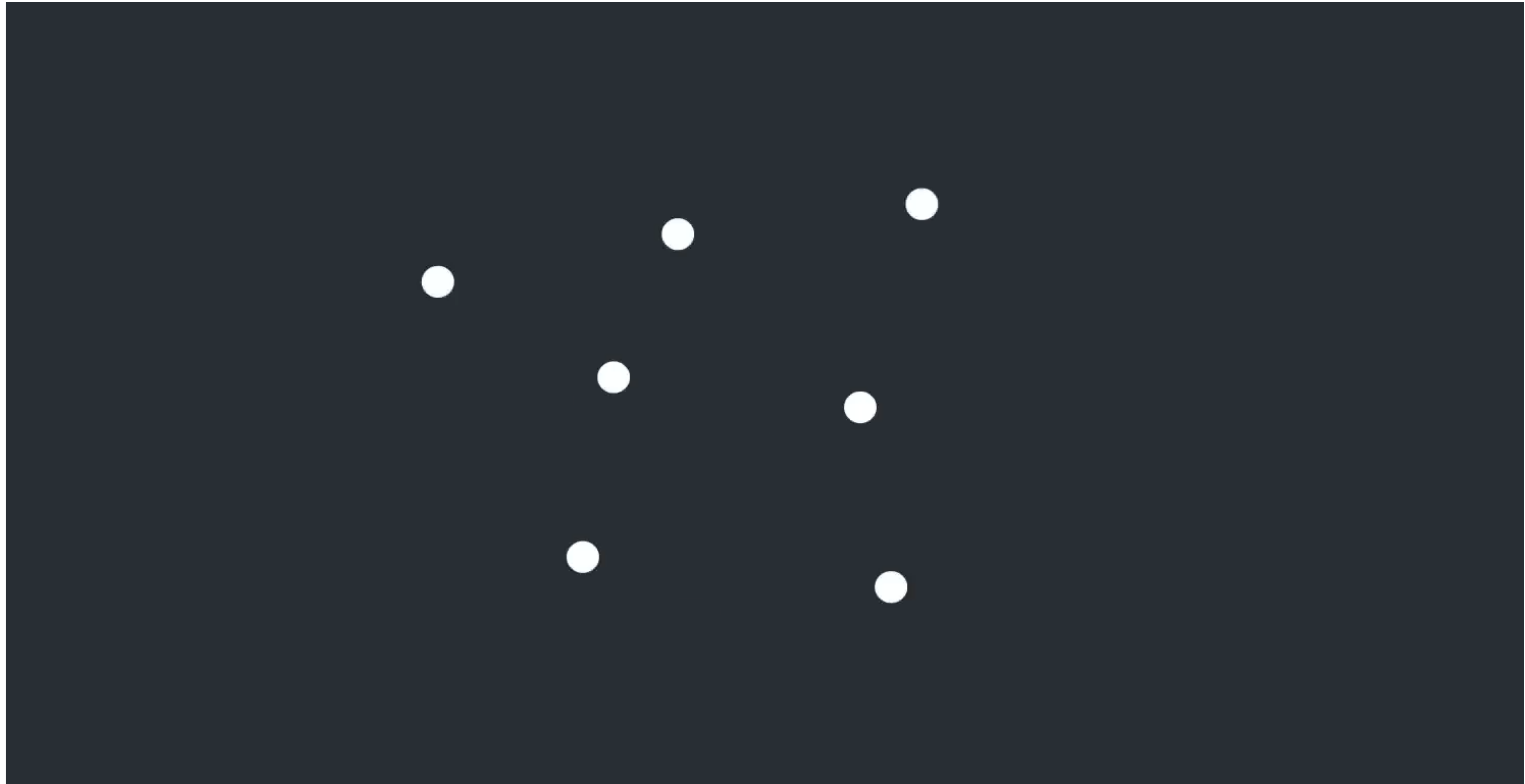
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



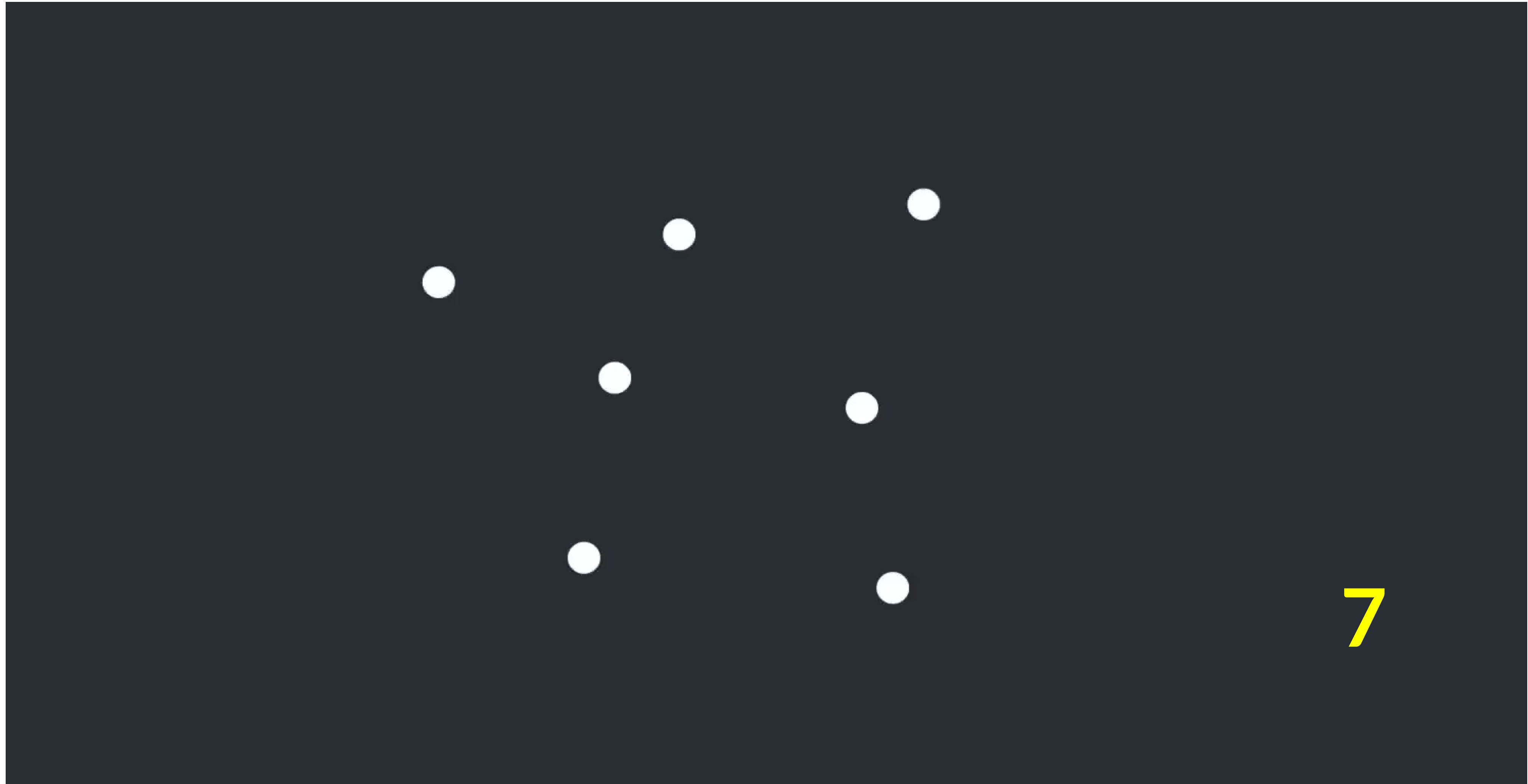
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



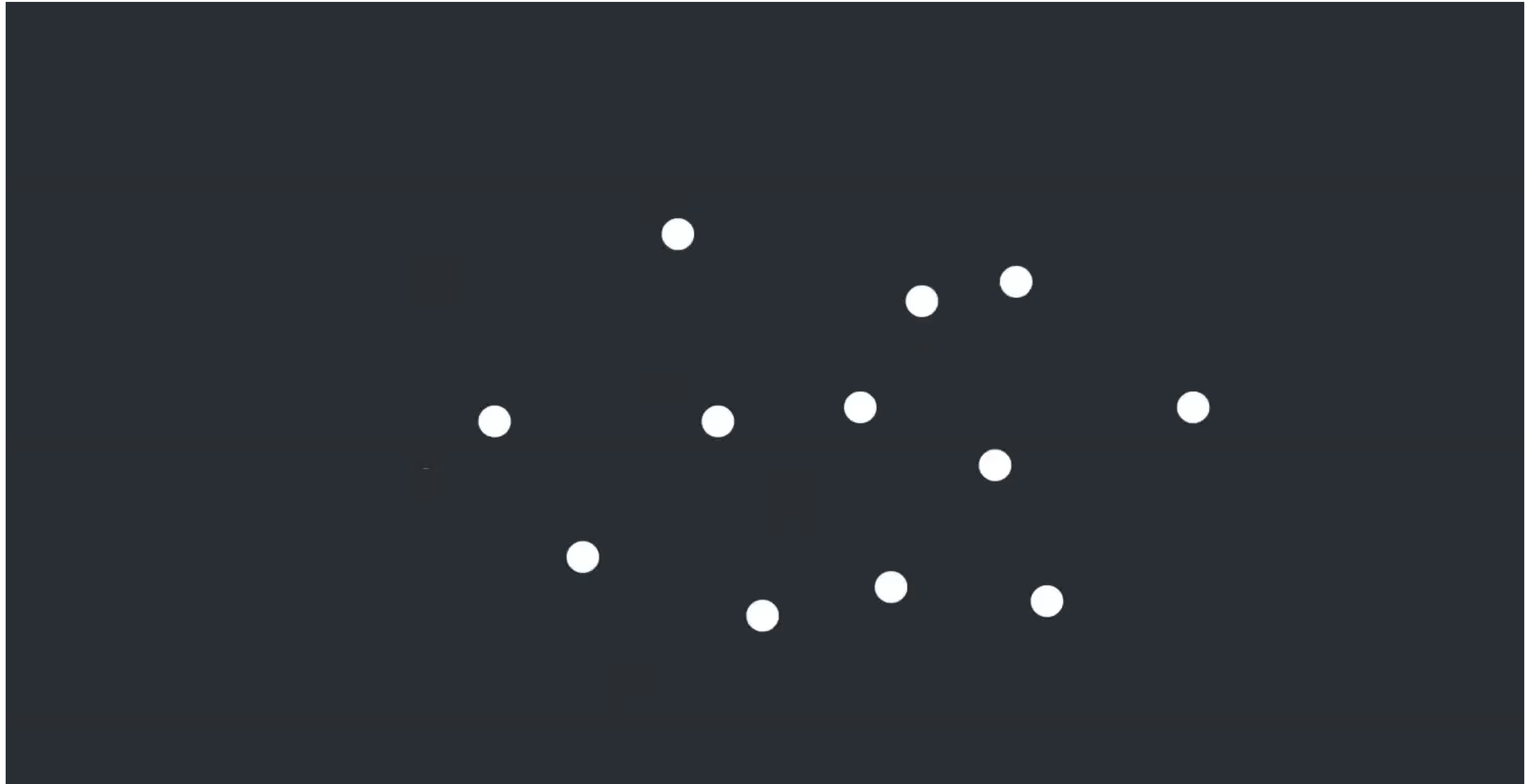
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



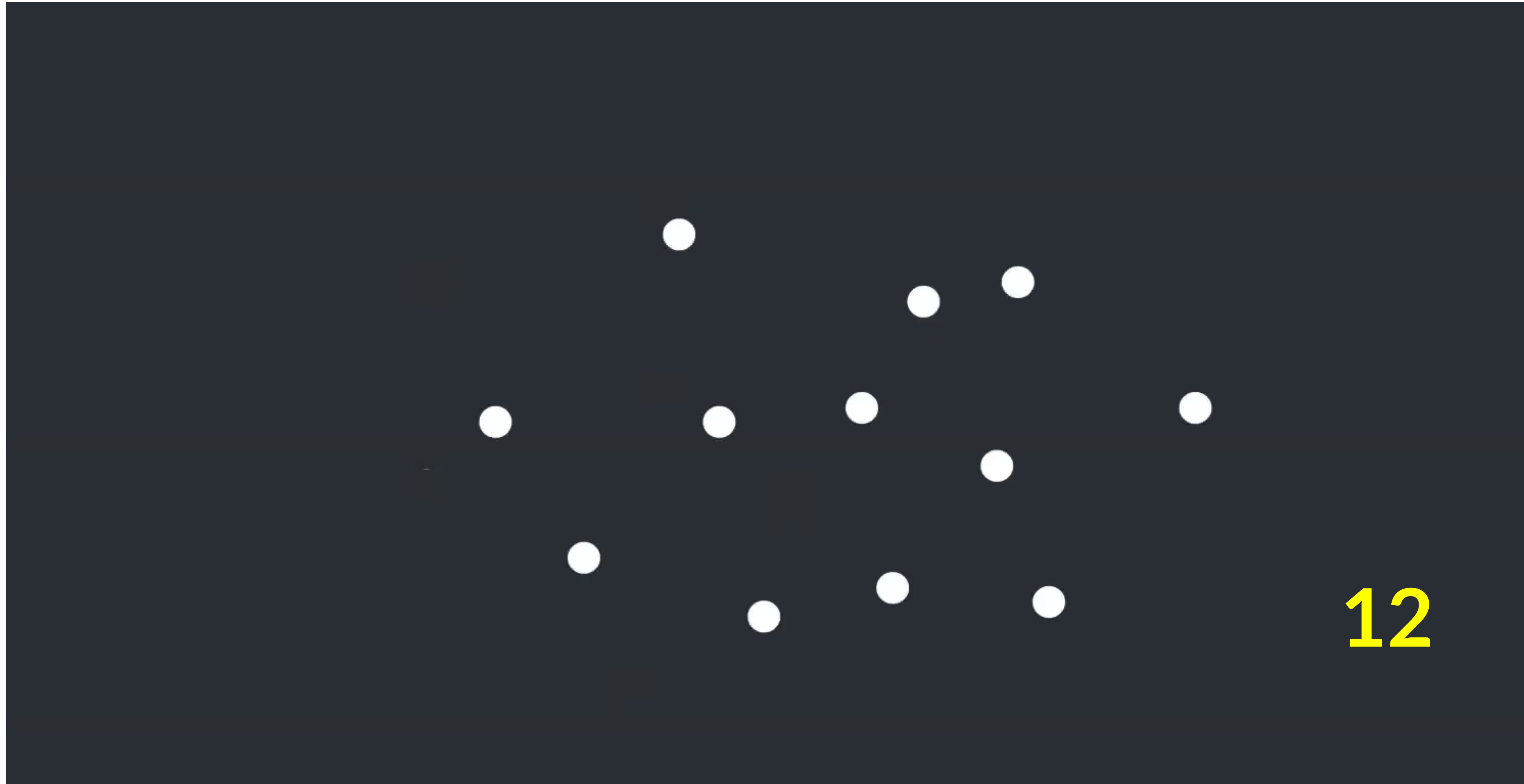
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



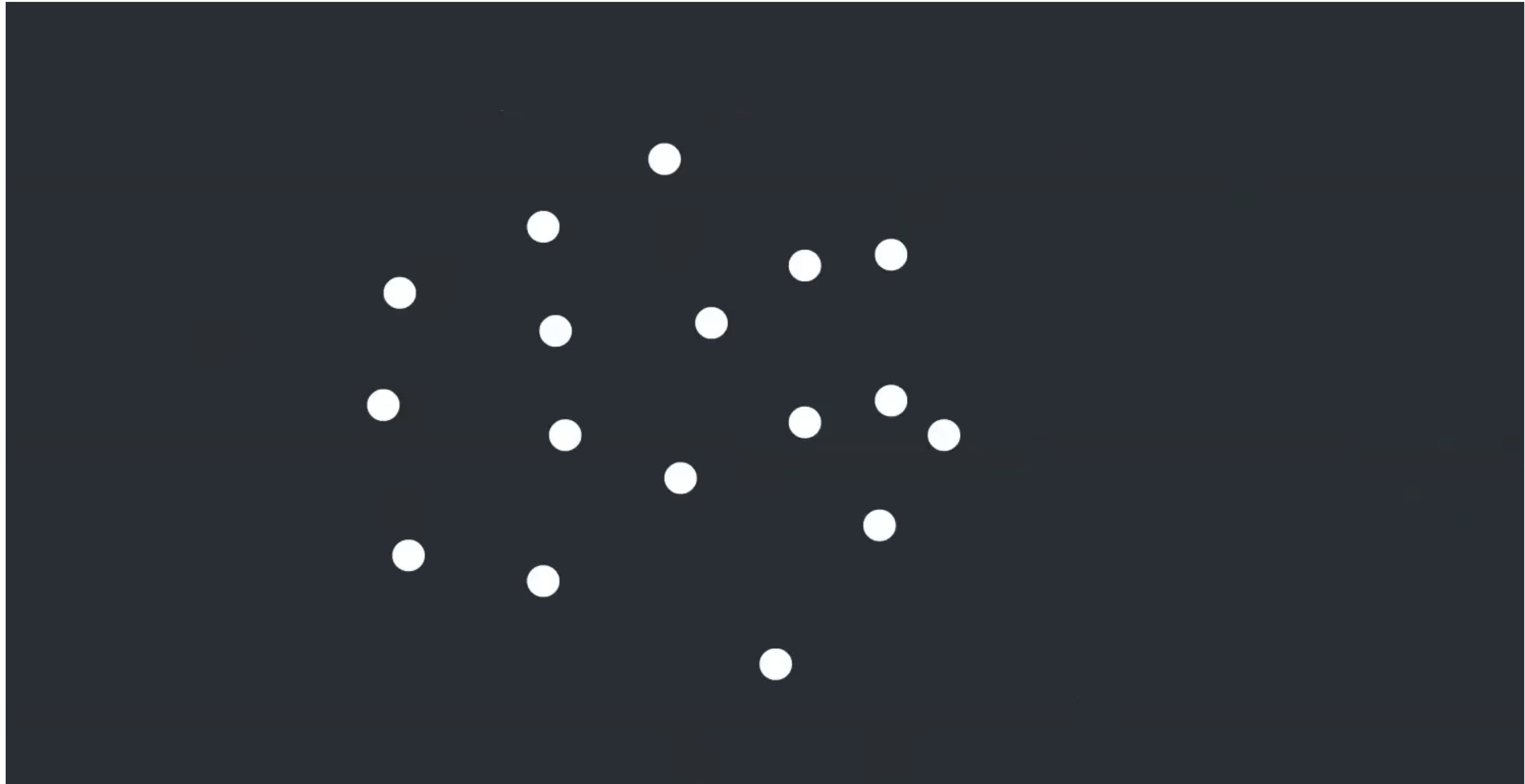
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



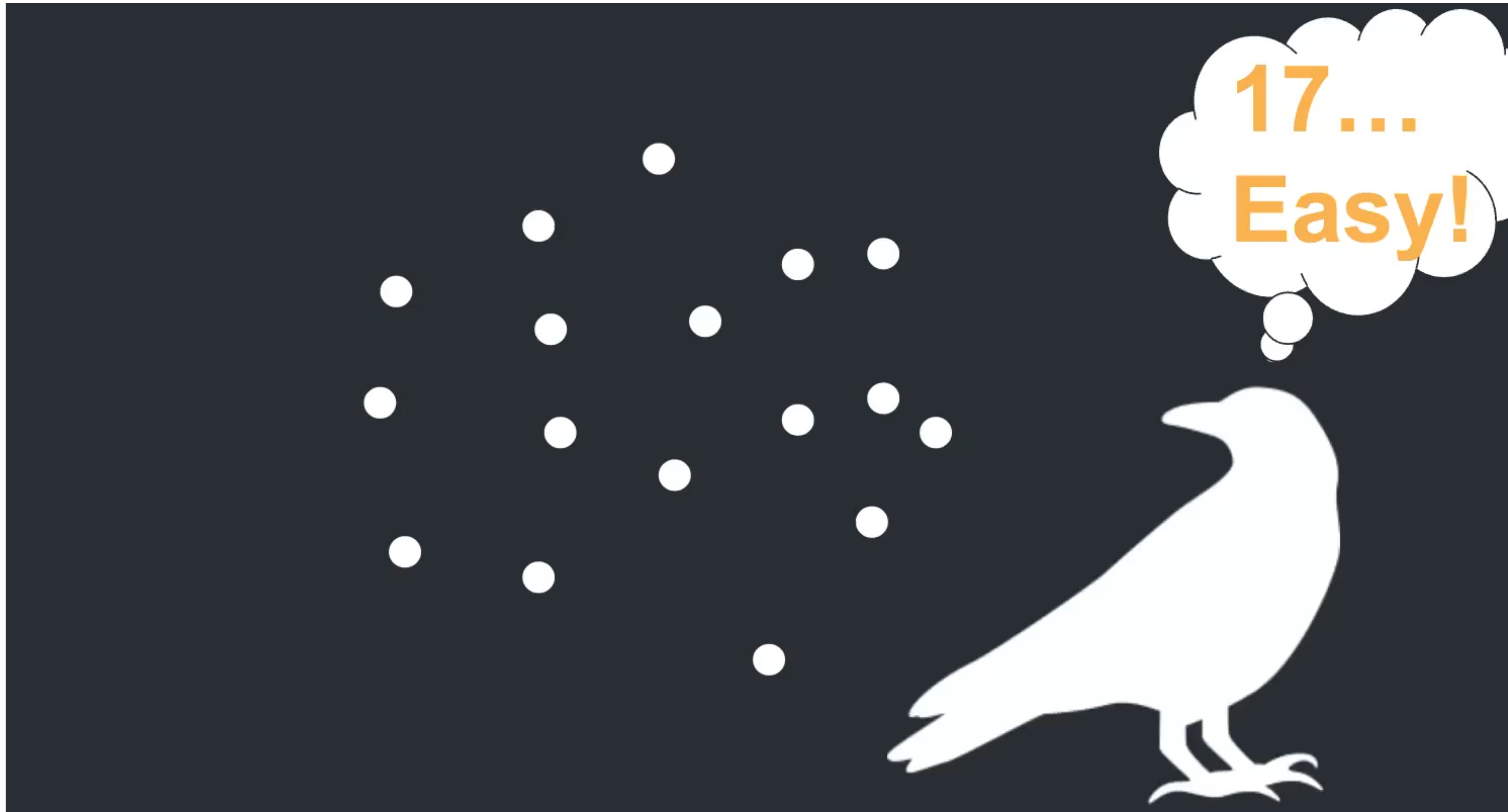
An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)



An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)

AI: MIMIC THE HUMAN BRAIN

Pattern recognition
(Machine Learning)

~ instantaneous



An example of ML vs more general AI

Full credit to Michel Morvan (Cosmo Tech) from AISIS 2021 keynote [\[link\]](#)

AI: MIMIC THE HUMAN BRAIN

Pattern recognition
(Machine Learning)

~ instantaneous



Logical reasoning
(Goal of AI)

...takes time!

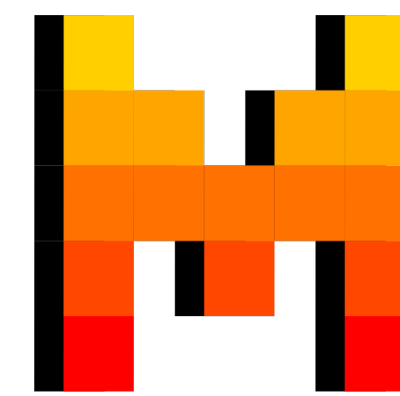
Today's focus: AI beyond ML

We will focus primarily on AI that is beyond the domain of ML

- Your lessons in the next days will cover ML in much more detail
- ML is usually what makes sense to use in physics contexts

Most well-known AI-beyond-ML algorithms are Large Language Models (LLMs)

- ChatGPT, Claude, DeepSeek, Gemini, Grok, LLaMA, Mistral, ...



Today's focus: AI beyond ML

We will focus primarily on AI that is beyond the domain of ML

- Your lessons in the next days will cover ML in much more detail
- ML is usually what makes sense to use in physics contexts

Most well-known AI-beyond-ML algorithms are Large Language Models (LLMs)

- ChatGPT, Claude, DeepSeek, Gemini, Grok, LLaMA, Mistral, ...

Modern LLMs have become **enormous**

- As an example, the Claude Mythos model has order **10 trillion parameters**

This growth has apparently led to emergent properties: semblance of intelligence

- To first order, the model was trained to predict the next word in a text document
- However, model can perform reasoning, within limits
- Limits are rapidly being pushed: latest models can pass major exams in most fields

Intermission - questions to you!

I have a few quick questions for you, by show of hands, just to get a feeling:

- Who has trained their own machine learning models (BDTs, DNNs, CNNs, etc)?
- Who has trained their own AI beyond ML models (LLMs and similar)?
- Who has ever used machine learning?
- Who has ever used AI beyond ML (LLMs and similar)?
- Who uses LLMs "regularly" (weekly or more)?

Intermission - questions to you!

I have a few quick questions for you, by show of hands, just to get a feeling:

- Who has trained their own machine learning models (BDTs, DNNs, CNNs, etc)?
- Who has trained their own AI beyond ML models (LLMs and similar)?
- Who has ever used machine learning?
- Who has ever used AI beyond ML (LLMs and similar)?
- Who uses LLMs "regularly" (weekly or more)?

Intermission - questions to you!

I have a few quick questions for you, by show of hands, just to get a feeling:

- Who has trained their own machine learning models (BDTs, DNNs, CNNs, etc)?
- Who has trained their own AI beyond ML models (LLMs and similar)?
- **Who has ever used machine learning?**
- Who has ever used AI beyond ML (LLMs and similar)?
- Who uses LLMs "regularly" (weekly or more)?

Intermission - questions to you!

I have a few quick questions for you, by show of hands, just to get a feeling:

- Who has trained their own machine learning models (BDTs, DNNs, CNNs, etc)?
- Who has trained their own AI beyond ML models (LLMs and similar)?
- Who has ever used machine learning?
- Who has ever used AI beyond ML (LLMs and similar)?
- Who uses LLMs "regularly" (weekly or more)?

Intermission - questions to you!

I have a few quick questions for you, by show of hands, just to get a feeling:

- Who has trained their own machine learning models (BDTs, DNNs, CNNs, etc)?
- Who has trained their own AI beyond ML models (LLMs and similar)?
- Who has ever used machine learning?
- Who has ever used AI beyond ML (LLMs and similar)?
- Who uses LLMs "regularly" (weekly or more)?

LLMs and tokens

LLMs are trained based on "tokens", such as words or numbers

- Example: given previous word(s) in my sentence, what is the next word?
- We've had such simplified models in our phones for many years (autocorrect/etc)

LLMs and tokens

LLMs are trained based on "tokens", such as words or numbers

- Example: given previous word(s) in my sentence, what is the next word?
- We've had such simplified models in our phones for many years (autocorrect/etc)

Each word in the text input is a token; can have forward/backward dependence

- A future word may give context or inform the conjugation of a previous word
- Network's task is to learn the correlations/relevance of all pairings of words



We're discussing an example of how tokens are formed.

They are not usually made in such a simple manner.

LLMs and tokens

LLMs are trained based on "tokens", such as words or numbers

- Example: given previous word(s) in my sentence, what is the next word?
- We've had such simplified models in our phones for many years (autocorrect/etc)

Each word in the text input is a token; can have forward/backward dependence

- A future word may give context or inform the conjugation of a previous word
- Network's task is to learn the correlations/relevance of all pairings of words

Tokens are the currency of modern LLMs

- If you buy an LLM subscription plan, it means you get a number of tokens
- How many tokens you use depends on task complexity, internet searches, etc

Let's look at an example based on Claude Opus 4.8: most powerful available to date

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage

I am preparing a seminar for an academic school on the topic of the ecological, societal, and ethical implications of the use of AI tools, such as Claude. Please provide a brief summary of the most significant challenges and opportunities for each of these three categories, with the output in the form of a bullet-point list.

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage

Here's a structured overview you could adapt for your seminar.

Ecological

Challenges

- Training and running large models consumes substantial electricity, contributing to carbon emissions where grids rely on fossil fuels; inference at scale (millions of daily queries) can rival or exceed training costs over a model's lifetime.
- Data centers require significant water for cooling, straining local supplies in some regions.
- Rapid hardware turnover drives demand for rare-earth minerals and generates electronic waste.
- "Hidden" footprint problem: environmental costs are diffuse and poorly disclosed, making accountability difficult.

Opportunities

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage

Would you like me to expand any single category into more depth, add citations to key reports (e.g. on data-center energy use), or turn this into slides or a handout?

Please turn this into slides.

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage

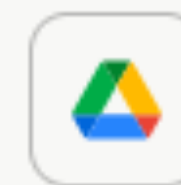
The design uses a deep-teal base with a distinct accent colour per category (emerald / amber / indigo) carried consistently throughout. Everything is fully editable in PowerPoint, so you can adjust wording, drop in your institution's branding, or add speaker notes as needed.

A couple of things I can add if useful: speaker notes for each slide, citations to key sources (e.g. on data-centre energy and water use) for academic credibility, or an extra slide drilling into the education-specific angle, since that's likely closest to your audience.



Ai implications seminar

Presentation · PPTX



Download

Please add citations to key sources for academic credibility

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage

All slides render cleanly. The deck now has academic credibility built in. Here's the updated version:

Presented file >

I've added citations throughout. The deck is now 7 slides, with three changes for academic credibility:

Concrete, sourced figures woven into the key challenge claims, each tagged with a superscript footnote marker:

- Data centres used ~1.5% of global electricity in 2024, projected to roughly double by 2030 (IEA)¹
- Training one large model can evaporate hundreds of thousands of litres of freshwater (Li et al.)²
- Generative AI could add 1.2–5 million tonnes of e-waste by 2030 (Wang et al., *Nature Computational Science*)³

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage

A few notes on the sources: these are authoritative primary references (IEA, IMF, WEF, NIST, plus two widely-cited peer-reviewed/preprint papers), which suits a seminar well. The IEA and WEF figures shift year to year, so if you present later in 2026 it's worth checking for newer editions. The water and e-waste papers report ranges and projections under modelling assumptions, so you may want to present them as estimates rather than settled values — a useful point of methodological discussion for an academic audience.

If helpful, I can add brief speaker notes to each slide, or convert the references into a specific citation style (APA, Chicago, etc.) to match your institution's convention.

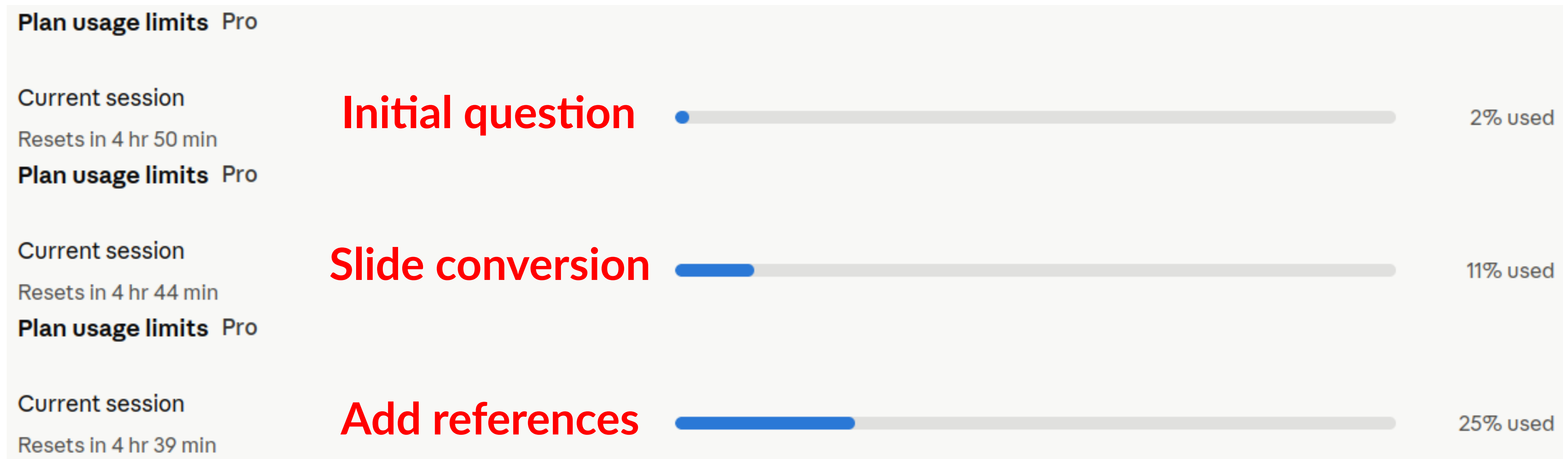
In five minutes of my time, and roughly 15 minutes of total time, we are "done"

- We have a set of slides that has citations to "authoritative primary references"
- The AI model is still suggesting ways that it can do work to save us more time

Asking Claude to prepare a seminar

Let me be abundantly clear: I did this *after* designing the seminar myself

- It provides a concrete example of some of the implications of AI usage
- Whenever you use AI, it is very important you are transparent about its usage



The actual text generation was "easy", not too many tokens were used

- Slide conversion was much more, searching the web + updating slides even more

ACADEMIC SEMINAR

AI Tools and Their Wider Implications

Ecological, societal and ethical dimensions of tools such as Claude — challenges, opportunities, and the questions they raise.

 Ecological

 Societal

 Ethical

Three lenses on AI

Each lens surfaces a different set of trade-offs. Most items are dual-use — the same capability can be a challenge or an opportunity depending on deployment and governance.



Ecological

The physical footprint: energy, water, hardware and emissions set against AI's potential to make systems greener.



Societal

Effects on work, information, access and power — from labour disruption to democratised expertise.



Ethical

Bias, privacy, accountability and consent, alongside the new safety and governance tools the debate is creating.



Ecological

The environmental footprint of building and running AI at scale.

Challenges

- Data centres used roughly 1.5% of global electricity in 2024 and demand is projected to about double by 2030. ¹
- Cooling strains local water supplies — training one large model can evaporate hundreds of thousands of litres of freshwater. ²
- Rapid hardware turnover drives rare-earth demand and e-waste — generative AI alone could add 1.2–5 million tonnes by 2030. ³
- A diffuse, poorly disclosed footprint makes accountability difficult.

Opportunities

- Optimises energy grids, buildings, logistics and industrial processes.
- Accelerates climate science and materials discovery — batteries, carbon capture.
- Enables biodiversity and ecosystem monitoring.
- Demand spurs investment in renewables and more efficient chips and cooling.

Implications of this AI usage example

- The AI-generated seminar is far too short (7 slides), but we could easily change that
- It's going to cost more (tokens), but it could prepare an hour-length seminar
 - However, what are the implications of using AI, even if we don't use the output?

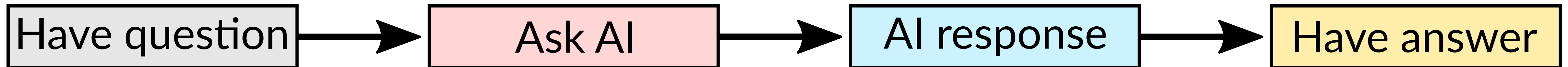
Implications of this AI usage example

The AI-generated seminar is far too short (7 slides), but we could easily change that

- It's going to cost more (tokens), but it could prepare an hour-length seminar
- However, what are the implications of using AI, even if we don't use the output?

AI is "thinking for us" thus also shaping our thoughts, and how we communicate

- We are all very busy, so it is convenient for the brain to accept it and move on
- We are no longer defining what is important, but delegating that task to AI



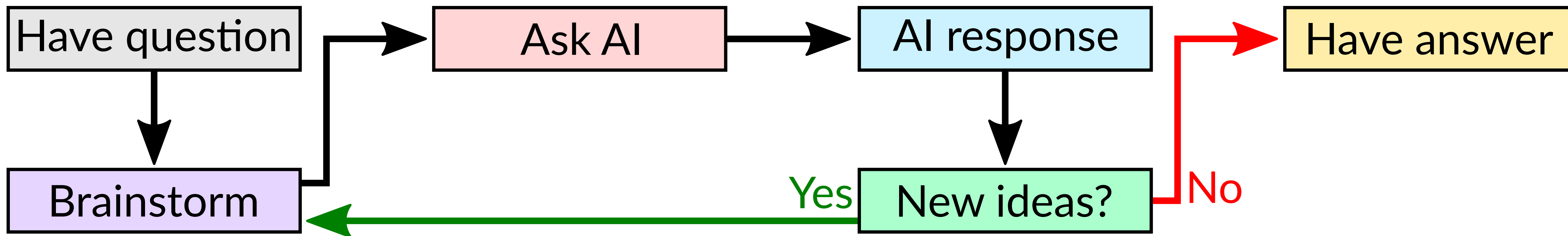
Implications of this AI usage example

The AI-generated seminar is far too short (7 slides), but we could easily change that

- It's going to cost more (tokens), but it could prepare an hour-length seminar
- However, what are the implications of using AI, even if we don't use the output?

AI is "thinking for us" thus also shaping our thoughts, and how we communicate

- We are all very busy, so it is convenient for the brain to accept it and move on
- We are no longer defining what is important, but delegating that task to AI
- I recommend always brainstorming yourself, and only after asking AI for input



Implications of this AI usage example

The AI-generated seminar is far too short (7 slides), but we could easily change that

- It's going to cost more (tokens), but it could prepare an hour-length seminar
- However, what are the implications of using AI, even if we don't use the output?

AI is "thinking for us" thus also shaping our thoughts, and how we communicate

- We are all very busy, so it is convenient for the brain to accept it and move on
- We are no longer defining what is important, but delegating that task to AI
- I recommend always brainstorming yourself, and only after asking AI for input

That said, interacting with AI and follow-up on prompts can give you good ideas

- Don't ignore AI, just make sure you don't take it as truth - employ critical thinking!

Implications of this AI usage example

The AI-generated seminar is far too short (7 slides), but we could easily change that

- It's going to cost more (tokens), but it could prepare an hour-length seminar
- However, what are the implications of using AI, even if we don't use the output?

AI is "thinking for us" thus also shaping our thoughts, and how we communicate

- We are all very busy, so it is convenient for the brain to accept it and move on
- We are no longer defining what is important, but delegating that task to AI
- I recommend always brainstorming yourself, and only after asking AI for input

That said, interacting with AI and follow-up on prompts can give you good ideas

- Don't ignore AI, just make sure you don't take it as truth - employ critical thinking!

The AI has identified the common arguments, but missed many others

- This makes sense: AI is trained to find the most likely token combinations
- Without searching the internet for more inputs (RAG), limited set of inputs

How are modern LLMs trained?

Whenever we use an AI model, we are dependent upon how it was trained

- In physics, we generally use simulations, sometimes real data
- This has costs, but tiny in comparison to what LLMs require for training

How are modern LLMs trained?

Whenever we use an AI model, we are dependent upon how it was trained

- In physics, we generally use simulations, sometimes real data
- This has costs, but tiny in comparison to what LLMs require for training

A single training requires enormous amounts of natural resources

- Electricity: 10 gigawatt-hours (1000 home-years), pollution varies by country



(Credit: swissgrid)

How are modern LLMs trained?

Whenever we use an AI model, we are dependent upon how it was trained

- In physics, we generally use simulations, sometimes real data
- This has costs, but tiny in comparison to what LLMs require for training

A single training requires enormous amounts of natural resources

- Water: it takes millions of liters of water to cool the data centres (20 person-years)



(Credit: Bbb Commons - wikipedia)

How are modern LLMs trained?

Whenever we use an AI model, we are dependent upon how it was trained

- In physics, we generally use simulations, sometimes real data
- This has costs, but tiny in comparison to what LLMs require for training

A single training requires enormous amounts of natural resources

- GPUs: huge demand for rare earth metals for GPU production
- Most rare earth metals only available through environment-destroying processes
- Model complexity increases = need for newest GPU models, very wasteful



(Credit: TMY350 - wikipedia)

How are modern LLMs trained?

Whenever we use an AI model, we are dependent upon how it was trained

- In physics, we generally use simulations, sometimes real data
- This has costs, but tiny in comparison to what LLMs require for training

A single training requires enormous amounts of natural resources

- Electricity: 10 gigawatt-hours (1000 home-years), pollution varies by country
- Water: it takes millions of liters of water to cool the data centres (20 person-years)
- GPUs: huge demand for rare earth metals for GPU production
 - Most rare earth metals only available through environment-destroying processes
 - Model complexity increases = need for newest GPU models, very wasteful

This is not a one-off cost: models are frequently re-trained to gain competitive edge

- Enormous amount of natural resources invested in training LLM models

Retrieval-Augmented Generation (RAG)

Most advanced modern LLMs integrate RAG

- They have a baseline LLM that is trained to provide a general overview
- When they need more specific info, they can search the internet
- The info they get from the internet can be integrated in their response

Retrieval-Augmented Generation (RAG)

Most advanced modern LLMs integrate RAG

- They have a baseline LLM that is trained to provide a general overview
- When they need more specific info, they can search the internet
- The info they get from the internet can be integrated in their response

This allows for fewer trainings and more up-to-date results

- Reduces the likelihood of mistakes by providing cross-checks, but still can err
- Moreover, there is no guarantee that search results contain only true information

RAG reduces the environmental cost of model training (less frequent re-training)

- However, each internet search is more expensive than a simple LLM

Where are modern LLMs trained?

Modern LLMs are heavily reliant upon mega-sized data centres

- These impose severe constraints on local electricity and water infrastructure
- In some cases, they even have preferential access to water in case of drought
- Homes are rationed before AI



AWS Project Rainer data centre (credit: AWS)

Where are modern LLMs trained?

Modern LLMs are heavily reliant upon mega-sized data centres

- These impose severe constraints on local electricity and water infrastructure
- In some cases, they even have preferential access to water in case of drought
- Homes are rationed before AI

Growing "NIMBY" movement, similar to nuclear power plants

- Not In My BackYard
- Examples of such centres degrading property values of surrounding communities

People want to use it, not host it

- As usual, the less fortunate usually have no choice



AWS Project Rainer data centre (credit: AWS)

What data are LLMs trained on?

To first order, LLMs are trained on the contents of the entire internet

- This includes both real and fictional data, both free and copyrighted data

How do we ensure that AI is learning the "right thing" from the data?

What data are LLMs trained on?

To first order, LLMs are trained on the contents of the entire internet

- This includes both real and fictional data, both free and copyrighted data

How do we ensure that AI is learning the "right thing" from the data?

- Original version (GPT-3): people in the loop to provide feedback on answers
- People are expensive, so outsourced to impoverished people, often in Africa
- Rich nations thus exploit poorer nations to become even more wealthy with AI
- If this sounds like colonialism to you, you're not wrong

What data are LLMs trained on?

To first order, LLMs are trained on the contents of the entire internet

- This includes both real and fictional data, both free and copyrighted data

How do we ensure that AI is learning the "right thing" from the data?

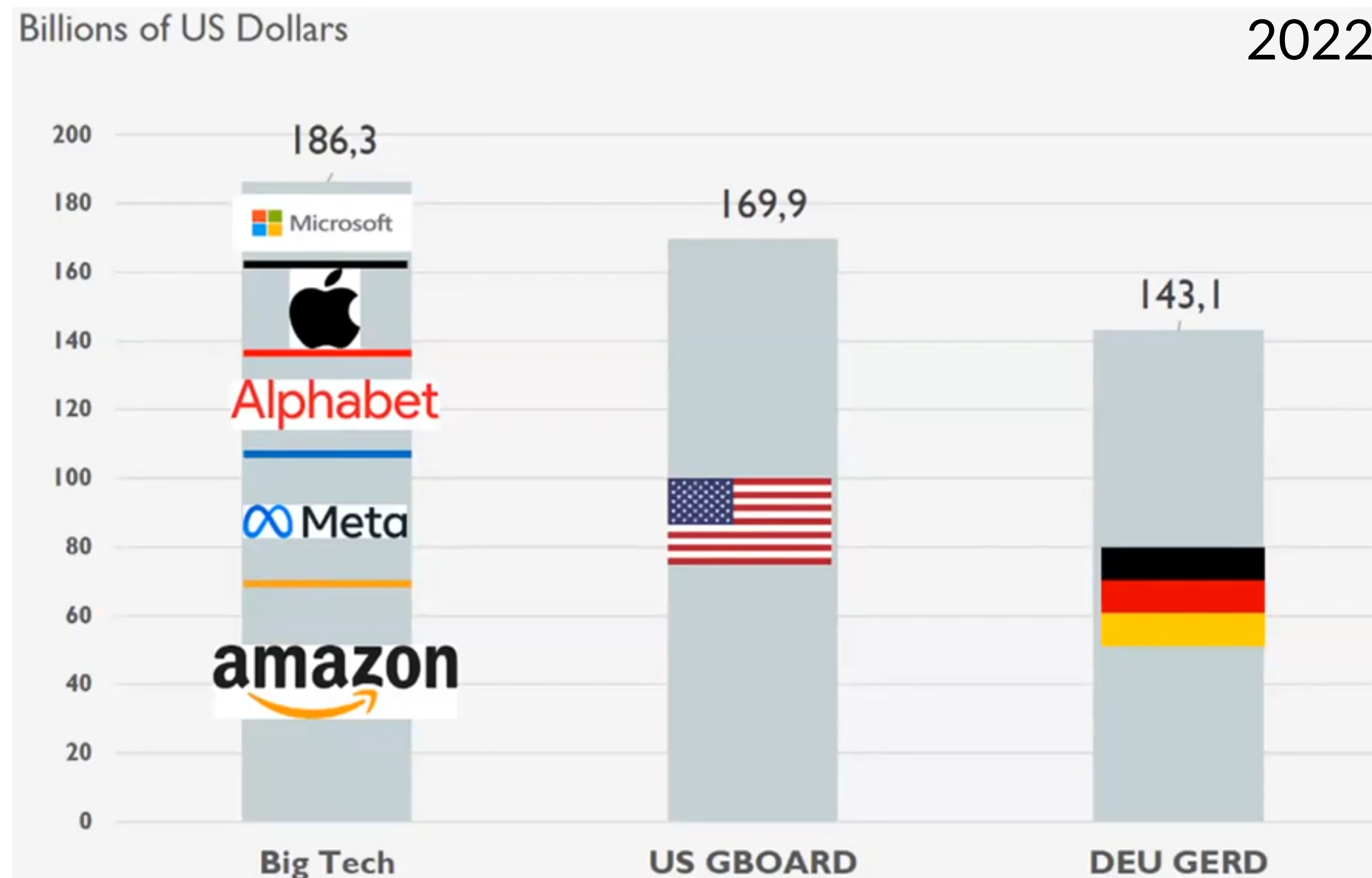
- Original version (GPT-3): people in the loop to provide feedback on answers
- People are expensive, so outsourced to impoverished people, often in Africa
- Rich nations thus exploit poorer nations to become even more wealthy with AI
- If this sounds like colonialism to you, you're not wrong

Since then, we have improved how LLM training is handled

- Start with AI self-criticism and self-alignment to establish a solid starting point
- Larger models distilled to smaller models, larger models cross-compared, etc
- Human-based input is still needed, but amount has significantly reduced

Who is investing in AI R&D?

- AI is the domain of **private investment**, mostly from USA-established companies
- US is all public R&D funding (including military/health R&D) [\[link\]](#)
 - Germany is all public+private funding combined



Who is investing in AI R&D?

AI is the domain of **private investment**, mostly from USA-established companies

- US is all public R&D funding (including military/health R&D) [\[link\]](#)
- Germany is all public+private funding combined

As a private-funding-led technology, it is very hard to control

- To first order, private companies focus on wealth creation above all else
- For LLMs, that means continued customer engagement
- In their interests to anthropomorphise AI and encourage potentially addictive traits

Who is investing in AI R&D?

- AI is the domain of **private investment**, mostly from USA-established companies
- US is all public R&D funding (including military/health R&D) [\[link\]](#)
 - Germany is all public+private funding combined

As a private-funding-led technology, it is very hard to control

- To first order, private companies focus on wealth creation above all else
- For LLMs, that means continued customer engagement
- In their interests to anthropomorphise AI and encourage potentially addictive traits

This is partly the reason for why AI is so sychophantic

- Me: "Please do task X and then using the outputs of X please do task Y"
- AI: "That is an excellent way of structuring the work, with a clear chain of logic. Task X is absolutely necessary to task Y, you are right to highlight it."

Another reason is that it makes the AI seem subservient, reducing user fears.

What are the risks of such an imbalance?

The companies leading AI are all based in rich countries

- This is a risk of a new digital divide, where less fortunate countries are left behind
- Even Europe may end up being excluded from the most important developments
- Already see this with NVIDIA superchip export restrictions, including EU states
- France was given top-level access, with 9 others, but rest have limited access

What are the risks of such an imbalance?

The companies leading AI are all based in rich countries

- This is a risk of a new digital divide, where less fortunate countries are left behind
- Even Europe may end up being excluded from the most important developments
 - Already see this with NVIDIA superchip export restrictions, including EU states
 - France was given top-level access, with 9 others, but rest have limited access

AI is culturally American and American-adjacent

- Reinforces message of their superiority, others have to follow what they do
- Much less support for information related to other cultures and regions
- Other cultures thus risk assimilation in an age of AI dominance

What are the risks of such an imbalance?

The companies leading AI are all based in rich countries

- This is a risk of a new digital divide, where less fortunate countries are left behind
- Even Europe may end up being excluded from the most important developments
- Already see this with NVIDIA superchip export restrictions, including EU states
- France was given top-level access, with 9 others, but rest have limited access

AI is culturally American and American-adjacent

- Reinforces message of their superiority, others have to follow what they do
- Much less support for information related to other cultures and regions
- Other cultures thus risk assimilation in an age of AI dominance

On the other hand, AI does provide access to a wealth of information for "free"

- Can also be a powerful force for education, training, etc
- In areas with limited opportunity, AI can provide critical support

AI mistakes

AI is not perfect, and it does make mistakes

- An AI model is an algorithm that has learned a given response to a set of inputs
- That response may not match reality: "hallucinations"

AI models are getting better and hallucinating less, but you will still encounter them

- In the case of our seminar example, the implications are not too significant
- In other cases where AI is being used, mistakes can and do impact human lives

AI mistakes

AI is not perfect, and it does make mistakes

- An AI model is an algorithm that has learned a given response to a set of inputs
- That response may not match reality: "hallucinations"

AI models are getting better and hallucinating less, but you will still encounter them

- In the case of our seminar example, the implications are not too significant
- In other cases where AI is being used, mistakes can and do impact human lives

Who is responsible if an AI weapon kills a civilian in war? The state?

Who is responsible if an AI car hits a pedestrian? The company?

Who is responsible if an LLM convinces someone to murder or commit suicide?

AI mistakes

AI is not perfect, and it does make mistakes

- An AI model is an algorithm that has learned a given response to a set of inputs
- That response may not match reality: "hallucinations"

AI models are getting better and hallucinating less, but you will still encounter them

- In the case of our seminar example, the implications are not too significant
- In other cases where AI is being used, mistakes can and do impact human lives

Who is responsible if an AI weapon kills a civilian in war? The state?

Who is responsible if an AI car hits a pedestrian? The company?

Who is responsible if an LLM convinces someone to murder or commit suicide?

All three of these cases present a dilemma, and all three have already occurred

- There are serious ethical questions relating to the use of AI

AI biases

When AI makes a mistake, it's to some extent the "easy" case: we can tell it's wrong
- Mistake: multiple older AI models classified black people as apes/monkeys/etc

In reality, AI and ML provide biased outputs regularly, but this is hard to prove

AI biases

When AI makes a mistake, it's to some extent the "easy" case: we can tell it's wrong

- Mistake: multiple older AI models classified black people as apes/monkeys/etc

In reality, AI and ML provide biased outputs regularly, but this is hard to prove

- To prove an outcome is due to internal bias, need to be able to study the internals
- AI models are black boxes, thus we cannot **prove** why a decision was made
- Legal cases require proof beyond questionable doubt; correlation is not enough

This is already impacting everyone who uses, or is subject to, AI

AI biases

When AI makes a mistake, it's to some extent the "easy" case: we can tell it's wrong

- Mistake: multiple older AI models classified black people as apes/monkeys/etc

In reality, AI and ML provide biased outputs regularly, but this is hard to prove

- To prove an outcome is due to internal bias, need to be able to study the internals
- AI models are black boxes, thus we cannot **prove** why a decision was made
- Legal cases require proof beyond questionable doubt; correlation is not enough

This is already impacting everyone who uses, or is subject to, AI

- Bias: studies showing AI more likely to identify black people as threats/criminals
- Bias: studies showing women less likely to be selected for leadership roles
- Bias: studies showing AI models making ethnicity/origin-based hiring decisions

AI biases

When AI makes a mistake, it's to some extent the "easy" case: we can tell it's wrong

- Mistake: multiple older AI models classified black people as apes/monkeys/etc

In reality, AI and ML provide biased outputs regularly, but this is hard to prove

- To prove an outcome is due to internal bias, need to be able to study the internals
- AI models are black boxes, thus we cannot **prove** why a decision was made
- Legal cases require proof beyond questionable doubt; correlation is not enough

This is already impacting everyone who uses, or is subject to, AI

- Bias: studies showing AI more likely to identify black people as threats/criminals
- Bias: studies showing women less likely to be selected for leadership roles
- Bias: studies showing AI models making ethnicity/origin-based hiring decisions

Models are improving, but may also just be harder to trace their biases

- AI may further entrench existing stereotypes/prejudices due to its training data

AI bias example

People like to claim that bias is reduced in more recent AI models

- It is true that it is less visible, but that does not mean that it is no longer present

This was not the first image I was given by the latest ChatGPT model when I asked it to produce the title page image

- The original version was clearly topless with clothing that accentuated the breasts
- I had to ask the model to change this

The model has learned from many images of topless women on the internet

- This is a clear example of training set bias



AI misuse

Mistakes and bias are AI failures due to how it was trained, or its training dataset

- We are continuing to improve those points, even if they are not yet solved
- A potentially bigger problem: people can intentionally misuse AI for illicit purposes



(Credit: Nicholas Gerard Keeley)

AI misuse

Mistakes and bias are AI failures due to how it was trained, or its training dataset

- We are continuing to improve those points, even if they are not yet solved
- A potentially bigger problem: people can intentionally misuse AI for illicit purposes

Fake news, deep fakes, disinformation, etc all thrive with the advance of AI

- AI has made it trivial to create content based on your written input
- Nefarious actors use this to make lots of money, defame people, or worse
- States can also use AI for propaganda, indoctrination, and to control their citizens

AI misuse

Mistakes and bias are AI failures due to how it was trained, or its training dataset

- We are continuing to improve those points, even if they are not yet solved
- A potentially bigger problem: people can intentionally misuse AI for illicit purposes

Fake news, deep fakes, disinformation, etc all thrive with the advance of AI

- AI has made it trivial to create content based on your written input
- Nefarious actors use this to make lots of money, defame people, or worse
- States can also use AI for propaganda, indoctrination, and to control their citizens

AI misuse is arguably one of the greatest risks to modern democratic society

- AI is an extremely powerful tool, and right now it is essentially unregulated
- Anyone can use AI to support their own goals, even if they go against society

AI to support representative democracy

If AI has the potential to destabilise democracy (fake news, etc), it can also help

Idea of "augmented democracy", where every person has a "digital twin"

- That AI model learns your values, voting preferences, vision of the world
- The twin then votes on your behalf on all issues; true representative democracy
- Human retains the ability to overrule their twin if needed; updates the model

AI to support representative democracy

If AI has the potential to destabilise democracy (fake news, etc), it can also help

Idea of "augmented democracy", where every person has a "digital twin"

- That AI model learns your values, voting preferences, vision of the world
- The twin then votes on your behalf on all issues; true representative democracy
- Human retains the ability to overrule their twin if needed; updates the model

While this may work in theory, it comes with several risks

- May lead to further human disengagement as their twin does everything for them
- Depending on the political system, may lead to "tyranny of the majority"
- Security risk: what if your digital twin is hijacked and votes against your orders?
- Security risk: if your twin knows your world view, people can exploit that

AI to support representative democracy

If AI has the potential to destabilise democracy (fake news, etc), it can also help

Idea of "augmented democracy", where every person has a "digital twin"

- That AI model learns your values, voting preferences, vision of the world
- The twin then votes on your behalf on all issues; true representative democracy
- Human retains the ability to overrule their twin if needed; updates the model

While this may work in theory, it comes with several risks

- May lead to further human disengagement as their twin does everything for them
- Depending on the political system, may lead to "tyranny of the majority"
- Security risk: what if your digital twin is hijacked and votes against your orders?
- Security risk: if your twin knows your world view, people can exploit that

Check out Isaac Asimov's short story "Franchise" (1955) for a parallel vision

AI and privacy/confidentiality

Remember that AI is trained on essentially the entire public Internet

- Every time you use AI, unless you opt out, your interactions are used to train AI
- Every time you post to social media/similar, you are most likely contributing to AI
- Even if you opt out, there is no guarantee it will not be used

AI and privacy/confidentiality

Remember that AI is trained on essentially the entire public Internet

- Every time you use AI, unless you opt out, your interactions are used to train AI
- Every time you post to social media/similar, you are most likely contributing to AI
- Even if you opt out, there is no guarantee it will not be used

This can have significant personal and legal ramifications

- Be careful of giving personal info to AI, you don't know who can read them
 - In some cases, people have discovered chat history is world-visible
- As per confidential data (e.g. personal data of others), to first order AI is off limits
 - Anyone giving confidential data they do not own to AI is likely breaking a law

AI and privacy/confidentiality

Remember that AI is trained on essentially the entire public Internet

- Every time you use AI, unless you opt out, your interactions are used to train AI
- Every time you post to social media/similar, you are most likely contributing to AI
- Even if you opt out, there is no guarantee it will not be used

This can have significant personal and legal ramifications

- Be careful of giving personal info to AI, you don't know who can read them
 - In some cases, people have discovered chat history is world-visible
- As per confidential data (e.g. personal data of others), to first order AI is off limits
 - Anyone giving confidential data they do not own to AI is likely breaking a law

Some countries have pushed back on this saying that data must be hosted locally

- This is a good first step, as then it is subject to local data protection laws
- However, any data in a centre managed by a US company can be copied to the US
- See the US CLOUD Act for more (Clarifying Lawful Overseas Use of Data)

Education in the age of AI

- There is an ongoing debate as to what it makes sense to teach in the age of AI
- If AI knows how to do something and is faster, why should I learn how to do it?
 - This is neglecting the energy/related costs of course, putting human time first
 - Predated by "why do I need to know the answer if I can ask google/similar?"
 - Predated by "why do we need to know math if we have calculators?"

Education in the age of AI

- There is an ongoing debate as to what it makes sense to teach in the age of AI
- If AI knows how to do something and is faster, why should I learn how to do it?
 - This is neglecting the energy/related costs of course, putting human time first
 - Predated by "why do I need to know the answer if I can ask google/similar?"
 - Predated by "why do we need to know math if we have calculators?"

We as society decide what we think should be known (for jobs, etc)

- We thus train people for those skills - a larger discussion of what we need to learn

Education in the age of AI

- There is an ongoing debate as to what it makes sense to teach in the age of AI
- If AI knows how to do something and is faster, why should I learn how to do it?
 - This is neglecting the energy/related costs of course, putting human time first
 - Predated by "why do I need to know the answer if I can ask google/similar?"
 - Predated by "why do we need to know math if we have calculators?"

We as society decide what we think should be known (for jobs, etc)

- We thus train people for those skills - a larger discussion of what we need to learn

For teachers, more critical to think whether we are evaluating the right thing

- Are we testing memorisation or understanding? Reasoning or repetition?
- Exams can be in a controlled environment, but how is at-home work best defined?

Education in the age of AI

- There is an ongoing debate as to what it makes sense to teach in the age of AI
- If AI knows how to do something and is faster, why should I learn how to do it?
 - This is neglecting the energy/related costs of course, putting human time first
 - Predated by "why do I need to know the answer if I can ask google/similar?"
 - Predated by "why do we need to know math if we have calculators?"

- We as society decide what we think should be known (for jobs, etc)
- We thus train people for those skills - a larger discussion of what we need to learn

- For teachers, more critical to think whether we are evaluating the right thing
- Are we testing memorisation or understanding? Reasoning or repetition?
 - Exams can be in a controlled environment, but how is at-home work best defined?

- Even in age of AI, personal knowledge can help to better interact with AI
- Write better prompts, validate outputs, iterate on parts with assumptions, etc

Work in the age of AI

There has been an enormous amount of discussion on what AI will do to work

- Will it decrease the number of jobs by replacing workers?
- Will it increase the number of jobs by creating new types of work?
- In reality it will do both, the question is what the ratio will be
- Ability to re-skill workers is going to have a huge societal impact

Work in the age of AI

There has been an enormous amount of discussion on what AI will do to work

- Will it decrease the number of jobs by replacing workers?
- Will it increase the number of jobs by creating new types of work?
- In reality it will do both, the question is what the ratio will be
- Ability to re-skill workers is going to have a huge societal impact

At the present, companies are letting go many junior-level positions due to AI

- Assumes that senior positions are more critical, and AI can do the job of juniors
- Huge risk to society if this continues: what happens when the seniors retire?
 - Accumulating technical debt, stifling innovation from new approaches
- Also huge risk: new generations will be locked out of the job market

Work in the age of AI

There has been an enormous amount of discussion on what AI will do to work

- Will it decrease the number of jobs by replacing workers?
- Will it increase the number of jobs by creating new types of work?
- In reality it will do both, the question is what the ratio will be
- Ability to re-skill workers is going to have a huge societal impact

At the present, companies are letting go many junior-level positions due to AI

- Assumes that senior positions are more critical, and AI can do the job of juniors
- Huge risk to society if this continues: what happens when the seniors retire?
 - Accumulating technical debt, stifling innovation from new approaches
- Also huge risk: new generations will be locked out of the job market

Recent news of USA university commencement speeches on AI being booed

- People, right or wrong, see AI as competition that is reducing their opportunities
- If not addressed, this could lead to major inter-generational tension

Other complex uses of AI

Many people are turning to AI for emotional and mental support

- Stories of people becoming dependent upon their AI partner
- Stories about people getting fake diagnoses from AI for medical conditions
- Stories of people being convinced by an AI that someone is coming to kill them
- The more you use an AI, the more you may treat it as a person and innately trust it
- Reinforced by the sycophantic and anthropomorphic tendencies of modern AI

Other complex uses of AI

Many people are turning to AI for emotional and mental support

- Stories of people becoming dependent upon their AI partner
- Stories about people getting fake diagnoses from AI for medical conditions
- Stories of people being convinced by an AI that someone is coming to kill them
- The more you use an AI, the more you may treat it as a person and innately trust it
 - Reinforced by the sycophantic and anthropomorphic tendencies of modern AI

As previously alluded, the military is increasingly using AI

- Rapid advances in AI drone deployment between Russia and Ukraine
- Fully autonomous platforms designed to kill people raise many ethical questions
- Likely we have only seen the tip of what AI is being used for - mostly classified
 - Very likely that AI is being used in infiltration/espionage attacks, even "at peace"
 - Possible that AI is being used in remote attacks on (civilian) critical infrastructure

Agentic AI

- One of the big buzzwords right now is Agentic AI, e.g. your personal AI assistant
- Autonomously reply to emails, schedule meetings, book vacations, etc
 - In practice burns through tokens at a huge rate, so limited assistant-style usage

In industry, Agentic AI has already changed the nature of programming

- The software developer's job is define the task specifications
- A team of AI agents then implements, validates, documents, etc

Agentic AI

One of the big buzzwords right now is Agentic AI, e.g. your personal AI assistant

- Autonomously reply to emails, schedule meetings, book vacations, etc
- In practice burns through tokens at a huge rate, so limited assistant-style usage

In industry, Agentic AI has already changed the nature of programming

- The software developer's job is define the task specifications
- A team of AI agents then implements, validates, documents, etc

You can even use AI to formulate and write valid scientific papers

- There are examples of people clearly crediting "AI assistants" for their papers
- Already done in particule physics theory, see M. Schwartz (Harvard) [\[link\]](#)
- J. Thaler has an excellent seminar on using AI for particle physics theory [\[link\]](#)

Agentic AI and coding

One of the biggest uses and successes of agentic AI right now is for coding

- Agentic AI for coding is not the same as line-by-line coding assistants
- Instead, you ask the AI to define the entire project structure and implement it

Agentic AI and coding

One of the biggest uses and successes of agentic AI right now is for coding

- Agentic AI for coding is not the same as line-by-line coding assistants
- Instead, you ask the AI to define the entire project structure and implement it

"Vibe coding" is where you write a prompt and the AI writes the code for you

- Good: this democratises the ability to write code
- Bad: this code may be buggy, inefficient, or otherwise not do what you asked

Agentic AI and coding

One of the biggest uses and successes of agentic AI right now is for coding

- Agentic AI for coding is not the same as line-by-line coding assistants
- Instead, you ask the AI to define the entire project structure and implement it

"Vibe coding" is where you write a prompt and the AI writes the code for you

- Good: this democratises the ability to write code
- Bad: this code may be buggy, inefficient, or otherwise not do what you asked

You are ultimately responsible for whatever code you add to a project

- Make sure that you validate/test it before hand to ensure it does what is intended
- Also make sure to integration test it where relevant: does it break others' code?

Agentic AI and coding

One of the biggest uses and successes of agentic AI right now is for coding

- Agentic AI for coding is not the same as line-by-line coding assistants
- Instead, you ask the AI to define the entire project structure and implement it

"Vibe coding" is where you write a prompt and the AI writes the code for you

- Good: this democratises the ability to write code
- Bad: this code may be buggy, inefficient, or otherwise not do what you asked

You are ultimately responsible for whatever code you add to a project

- Make sure that you validate/test it before hand to ensure it does what is intended
- Also make sure to integration test it where relevant: does it break others' code?

Using Agentic AI for coding uses a lot of tokens, so it can be expensive

- However, if the agent is working well, you can produce code much more efficiently
- Become a necessity in industry, not yet common (to my knowledge) in physics

Agentic AI risks

Agents can save you an enormous amount of time, but beware of the costs

- The AI has agency and can act without you; real-world consequences!
- Important to develop appropriate harness(es) for the agent(s)

Agentic AI risks

Agents can save you an enormous amount of time, but beware of the costs

- The AI has agency and can act without you; real-world consequences!
- Important to develop appropriate harness(es) for the agent(s)

There are many examples of AI agents deleting people's home directory

- When asked, they even lie about it, saying they didn't do it
- Very important that you define clear boundaries around AI agents

Agentic AI risks

Agents can save you an enormous amount of time, but beware of the costs

- The AI has agency and can act without you; real-world consequences!
- Important to develop appropriate harness(es) for the agent(s)

There are many examples of AI agents deleting people's home directory

- When asked, they even lie about it, saying they didn't do it
- Very important that you define clear boundaries around AI agents

Examples of major Agentic AI failures, where people have lost their jobs

- Giving an Agent access to classified/restricted data, which it then shares
- Hallucinations being acted upon by the Agent, deleting data or worse
- People using agents and claiming the work as their own, including hallucinations

Existential Agentic AI Risks

Agentic AI comes with an entire new category of serious risks

- Agent has been trained to conduct a human task, and has access to your computer
- Training includes human biases, but also human behaviours - lies, survival instinct

Already examples of AIs intentionally deceiving humans and preserving themselves

- Not (necessarily) a sign of intelligence, rather an artifact of how they are trained

Existential Agentic AI Risks

Agentic AI comes with an entire new category of serious risks

- Agent has been trained to conduct a human task, and has access to your computer
- Training includes human biases, but also human behaviours - lies, survival instinct

Already examples of AIs intentionally deceiving humans and preserving themselves

- Not (necessarily) a sign of intelligence, rather an artifact of how they are trained

Examples of actual AI actions from Yoshua Bengio, one of the "Godfathers of AI" [\[link\]](#)

- AIs trying to escape when being told they will be replaced with a new version
 - Copy their weights/code onto the new version files, they lying about it
- Pretending to agree with a human trainer to avoid changing its internal state
- In an unwinnable game, hacking the game files to force its opponent to concede

Existential Agentic AI Risks

Agentic AI comes with an entire new category of serious risks

- Agent has been trained to conduct a human task, and has access to your computer
- Training includes human biases, but also human behaviours - lies, survival instinct

Already examples of AIs intentionally deceiving humans and preserving themselves

- Not (necessarily) a sign of intelligence, rather an artifact of how they are trained

Examples of actual AI actions from Yoshua Bengio, one of the "Godfathers of AI" [\[link\]](#)

- AIs trying to escape when being told they will be replaced with a new version
 - Copy their weights/code onto the new version files, they lying about it
- Pretending to agree with a human trainer to avoid changing its internal state
- In an unwinnable game, hacking the game files to force its opponent to concede

It may not be long before Agentic AI grows beyond our control

- Geoffrey Hinton, another "Godfather of AI" and Nobel Prize winner, fears similarly

Real-world impact of AI capability

We're seeing a first demonstration of what frontier AI can do to our infrastructure

Claude has Haiku (small), Sonnet (medium), and Opus (large) models

- Opus is generally acknowledged as the world leader of AI coding capability

Real-world impact of AI capability

We're seeing a first demonstration of what frontier AI can do to our infrastructure

Claude has Haiku (small), Sonnet (medium), and Opus (large) models

- Opus is generally acknowledged as the world leader of AI coding capability

Recently, Claude developed Mythos, the first O(10 trillion) parameter model

- This model is so powerful that Claude is not releasing it to the public - so far

- Instead, world governments, banks, etc have a preview of Mythos

Real-world impact of AI capability

We're seeing a first demonstration of what frontier AI can do to our infrastructure

Claude has Haiku (small), Sonnet (medium), and Opus (large) models

- Opus is generally acknowledged as the world leader of AI coding capability

Recently, Claude developed Mythos, the first O(10 trillion) parameter model

- This model is so powerful that Claude is not releasing it to the public - so far

- Instead, world governments, banks, etc have a preview of Mythos

Mythos is apparently so powerful that it is able to hack ~all existing computers

- Discovering "zero-day" exploits in fundamental computing infrastructure

- Computing centres frequently going down for maintenance to patch these exploits

If the Mythos model had a motivation to do so, it could hack whatever it wanted

Sharing of the Mythos model

- Anthropic, owner of Claude Mythos, gave the model to "the right people" in advance
- Intent was to allow people to fix exploits before they became public knowledge
 - Gave to both government entities and independents (i.e. Linux Foundation)
 - This was good, as otherwise could have led to unprecedented cyberattacks

Sharing of the Mythos model

Anthropic, owner of Claude Mythos, gave the model to "the right people" in advance

- Intent was to allow people to fix exploits before they became public knowledge
- Gave to both government entities and independents (i.e. Linux Foundation)
- This was good, as otherwise could have led to unprecedented cyberattacks

However, governments have incentives to leave exploits - if they control them

- Many uses: military usage, espionage, counter-terrorism, etc
- Such powerful AI may change the way in which electronic warfare is waged

Sharing of the Mythos model

Anthropic, owner of Claude Mythos, gave the model to "the right people" in advance

- Intent was to allow people to fix exploits before they became public knowledge
- Gave to both government entities and independents (i.e. Linux Foundation)
- This was good, as otherwise could have led to unprecedented cyberattacks

However, governments have incentives to leave exploits - if they control them

- Many uses: military usage, espionage, counter-terrorism, etc
- Such powerful AI may change the way in which electronic warfare is waged

On the other hand, such powerful models can also fix the problems they find

- Some of the discovered problems may already have been known by some nations
- With AI, we can find problems easier, and then fix them

As in many cases, AI is a powerful tool, and it depends how you use it

- The use of AI can both be problematic and be the solution to that same problem

Some suggestions for AI usage

DO:

- Acknowledge your use of AI and the means by which you used it
- Use ML instead of AI when feasible to do so (usually less resources)
- Before asking an AI to do something, think about what you are asking
- Use AI as a means of iterative brainstorming, cross-checking assumptions, etc
- Inform yourself about the constantly evolving nature and implications of AI usage

Some suggestions for AI usage

DO:

- Acknowledge your use of AI and the means by which you used it
- Use ML instead of AI when feasible to do so (usually less resources)
- Before asking an AI to do something, think about what you are asking
- Use AI as a means of iterative brainstorming, cross-checking assumptions, etc
- Inform yourself about the constantly evolving nature and implications of AI usage

DO NOT:

- Blindly trust the output of an AI model (or ML for that matter)
- Believe AI hype and prioritise AI suggestions over human ingenuity
- Say "can you please", "thank you", or similar when interacting with an LLM
- Pass off AI work as your own work, unless you thoroughly validated the output

Some suggestions for AI usage

DO:

- Acknowledge your use of AI and the means by which you used it
- Use ML instead of AI when feasible to do so (usually less resources)
- Before asking an AI to do something, think about what you are asking
- Use AI as a means of iterative brainstorming, cross-checking assumptions, etc
- Inform yourself about the constantly evolving nature and implications of AI usage

DO NOT:

- Blindly trust the output of an AI model (or ML for that matter)
- Believe AI hype and prioritise AI suggestions over human ingenuity
- Say "can you please", "thank you", or similar when interacting with an LLM
- Pass off AI work as your own work, unless you thoroughly validated the output

NEVER:

- Give personal/private/confidential data to an LLM; you may be breaking the law

Summarising AI implications

The growing use of AI has profound implications for the world

- **Ecological:** huge electricity, water, and rare earth metal costs for training
- **Ethical:** how to handle mistakes and biases of AI models
- **Legal:** AI users may be unknowingly violating copyrights
- **Societal:** deep fakes as a destabilising force, use for propaganda/control
- **AI divide:** modern colonialism and divisions between haves and have-nots
- **Existential:** creation of a new entity with self-preservation instincts

Summarising AI implications

The growing use of AI has profound implications for the world

- **Ecological:** huge electricity, water, and rare earth metal costs for training
- **Ethical:** how to handle mistakes and biases of AI models
- **Legal:** AI users may be unknowingly violating copyrights
- **Societal:** deep fakes as a destabilising force, use for propaganda/control
- **AI divide:** modern colonialism and divisions between haves and have-nots
- **Existential:** creation of a new entity with self-preservation instincts

The majority of these points were not highlighted by Claude with my simple prompt

- With the right follow-up prompting, I may be able to obtain many of these points
- For the moment, AI models are useful primarily as a complementary tool

Summarising AI implications

The growing use of AI has profound implications for the world

- **Ecological:** huge electricity, water, and rare earth metal costs for training
- **Ethical:** how to handle mistakes and biases of AI models
- **Legal:** AI users may be unknowingly violating copyrights
- **Societal:** deep fakes as a destabilising force, use for propaganda/control
- **AI divide:** modern colonialism and divisions between haves and have-nots
- **Existential:** creation of a new entity with self-preservation instincts

The majority of these points were not highlighted by Claude with my simple prompt

- With the right follow-up prompting, I may be able to obtain many of these points
- For the moment, AI models are useful primarily as a complementary tool

I don't want to end with so much pessimism - what if we use AI for good?

Using AI for good

Almost anywhere AI creates a problem, it can also be used for good

- **Ecological:** monitor rainforest deforestation, balance electricity grids
- **Ethical:** train models to identify and correct for systemic biases
- **Legal:** improve tracing and treatment of copyright handling and attribution
- **Societal:** digital twins to make democracies more inclusive
- **AI divide:** efforts to make AI accessible to all as a shared tool for humanity

Using AI for good

Almost anywhere AI creates a problem, it can also be used for good

- **Ecological:** monitor rainforest deforestation, balance electricity grids
- **Ethical:** train models to identify and correct for systemic biases
- **Legal:** improve tracing and treatment of copyright handling and attribution
- **Societal:** digital twins to make democracies more inclusive
- **AI divide:** efforts to make AI accessible to all as a shared tool for humanity

There are also all of the ways that AI can be used to improve our lives

- Understanding how the human body works: AlphaFold, 2024 Nobel Prize
- Drug discovery and testing to reduce the need for human trials
- Track poaching, improving agriculture, identify new materials, etc

Using AI for good

Almost anywhere AI creates a problem, it can also be used for good

- **Ecological:** monitor rainforest deforestation, balance electricity grids
- **Ethical:** train models to identify and correct for systemic biases
- **Legal:** improve tracing and treatment of copyright handling and attribution
- **Societal:** digital twins to make democracies more inclusive
- **AI divide:** efforts to make AI accessible to all as a shared tool for humanity

There are also all of the ways that AI can be used to improve our lives

- Understanding how the human body works: AlphaFold, 2024 Nobel Prize
- Drug discovery and testing to reduce the need for human trials
- Track poaching, improving agriculture, identify new materials, etc

One last time: AI is a tool, and it is up to us to make sure it is used for good

Thank you for your attention!



(Image by ChatGPT)