Gerry Chng
Email: Z2VycnkuY2huZ0BnbWFpbC5jb20=

Algorithmic weaknesses

Is the Cybersecurity workforce ready for the future?

AGENDA

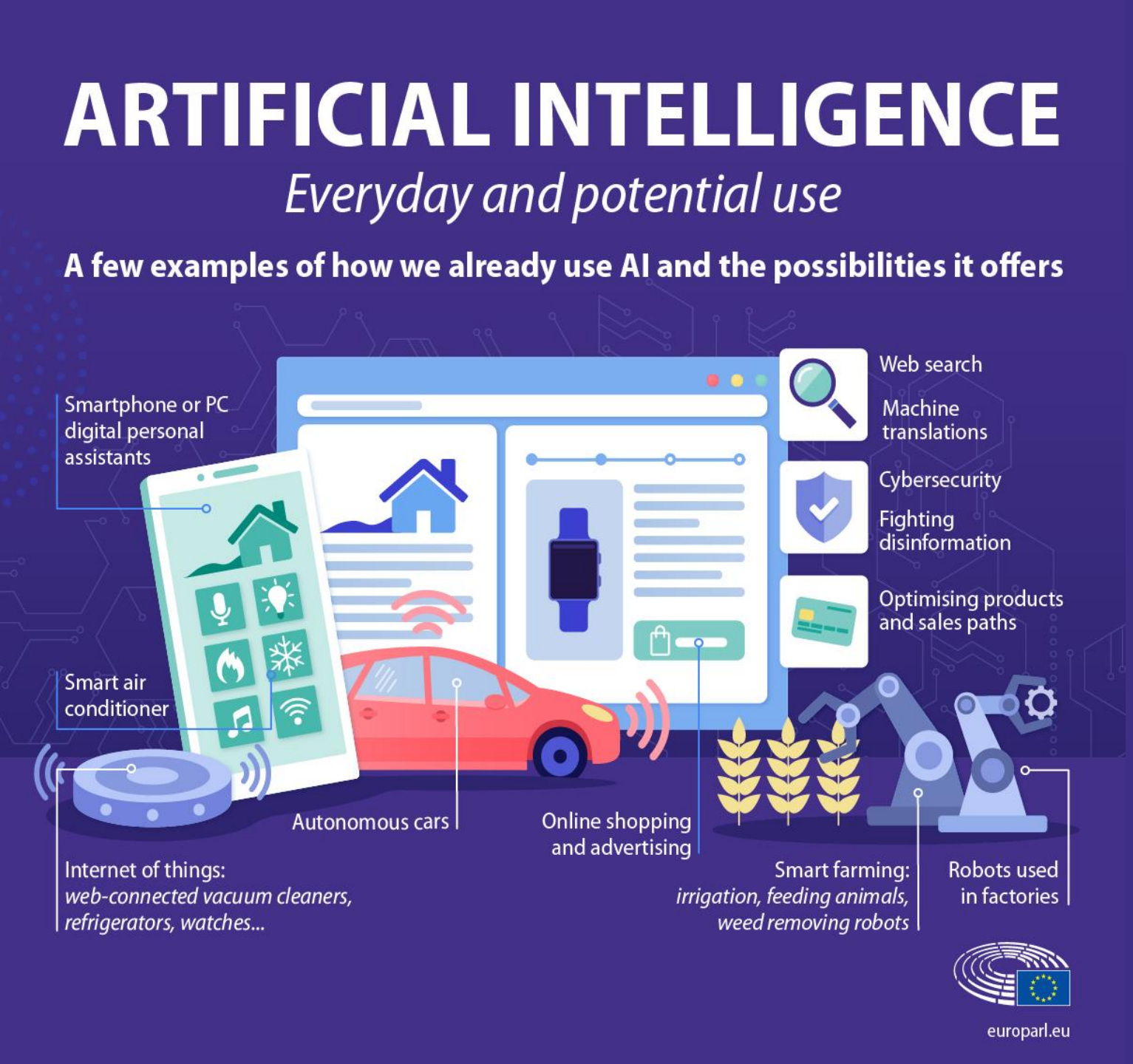
How have algorithms **influenced your day?**

But **what exactly** is AI?

What can go **wrong**?

What **skills** are required for the **future**?

How have algorithms influenced your day?

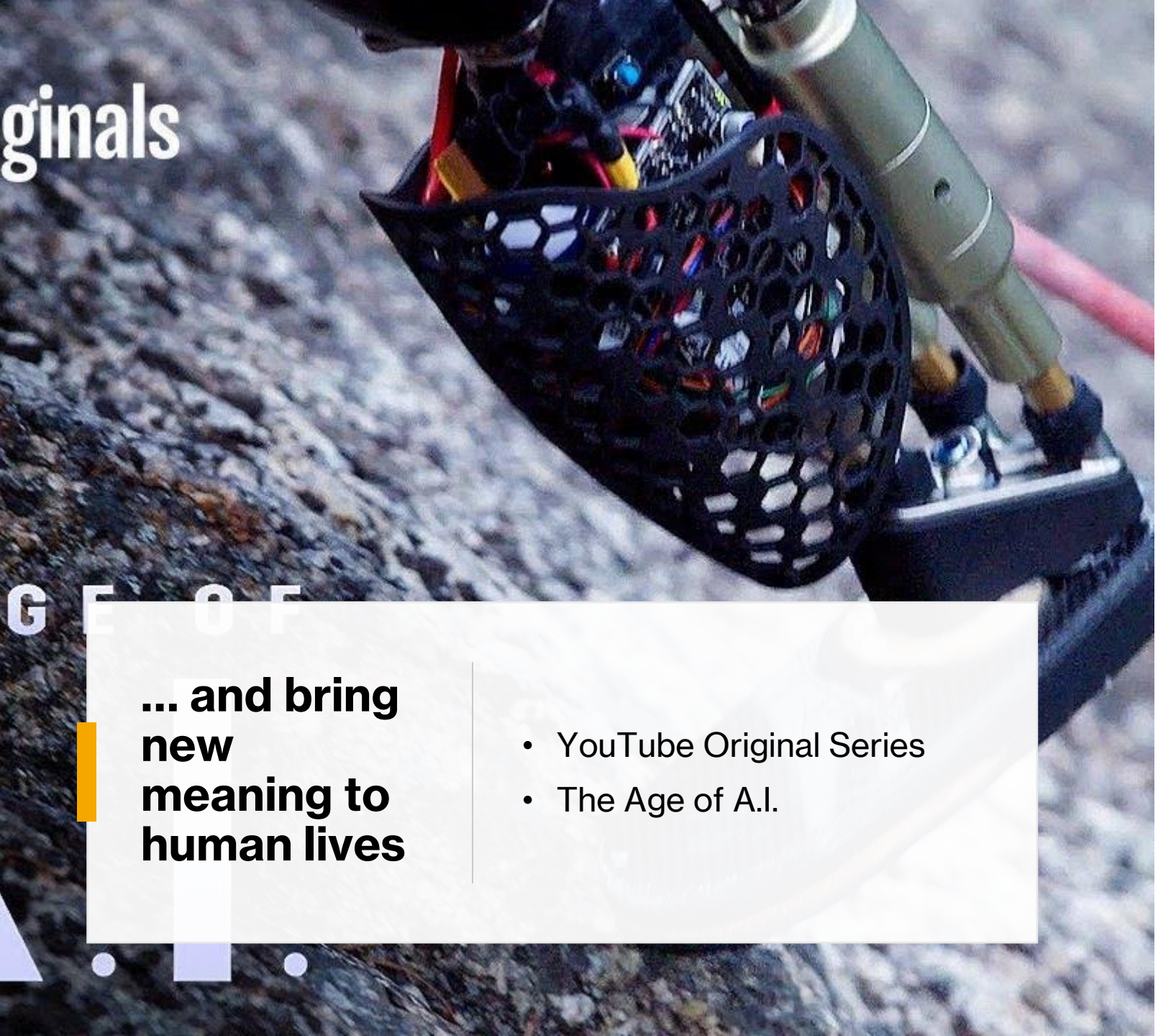


Increasingly, our

daily activities, interactions, and decisions

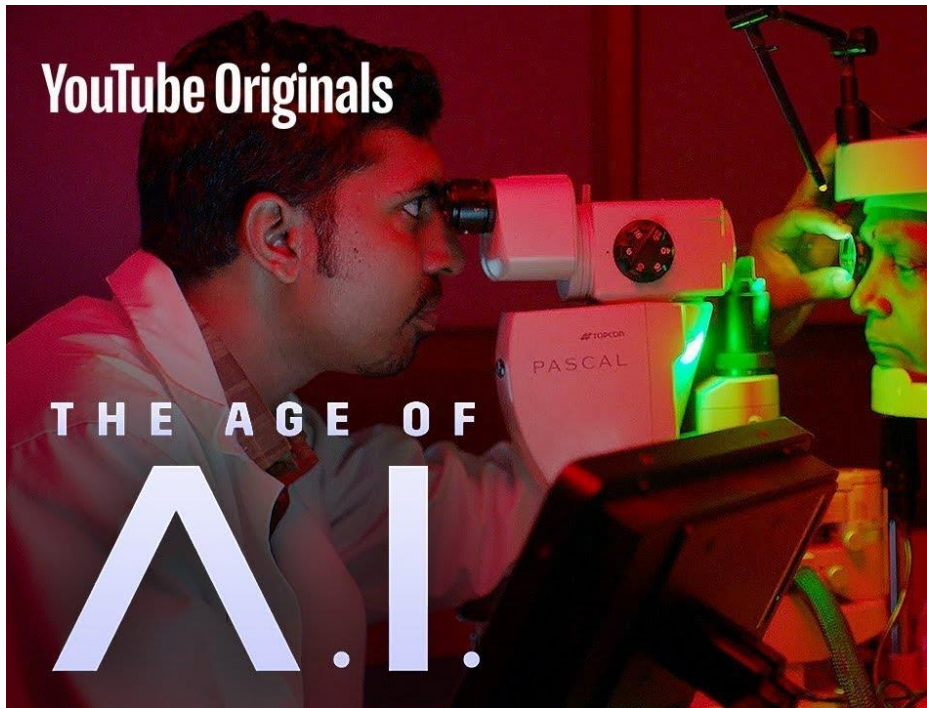
are nudged along by algorithms

It allows businesses to optimize their operations ...



... and bring new meaning to human lives

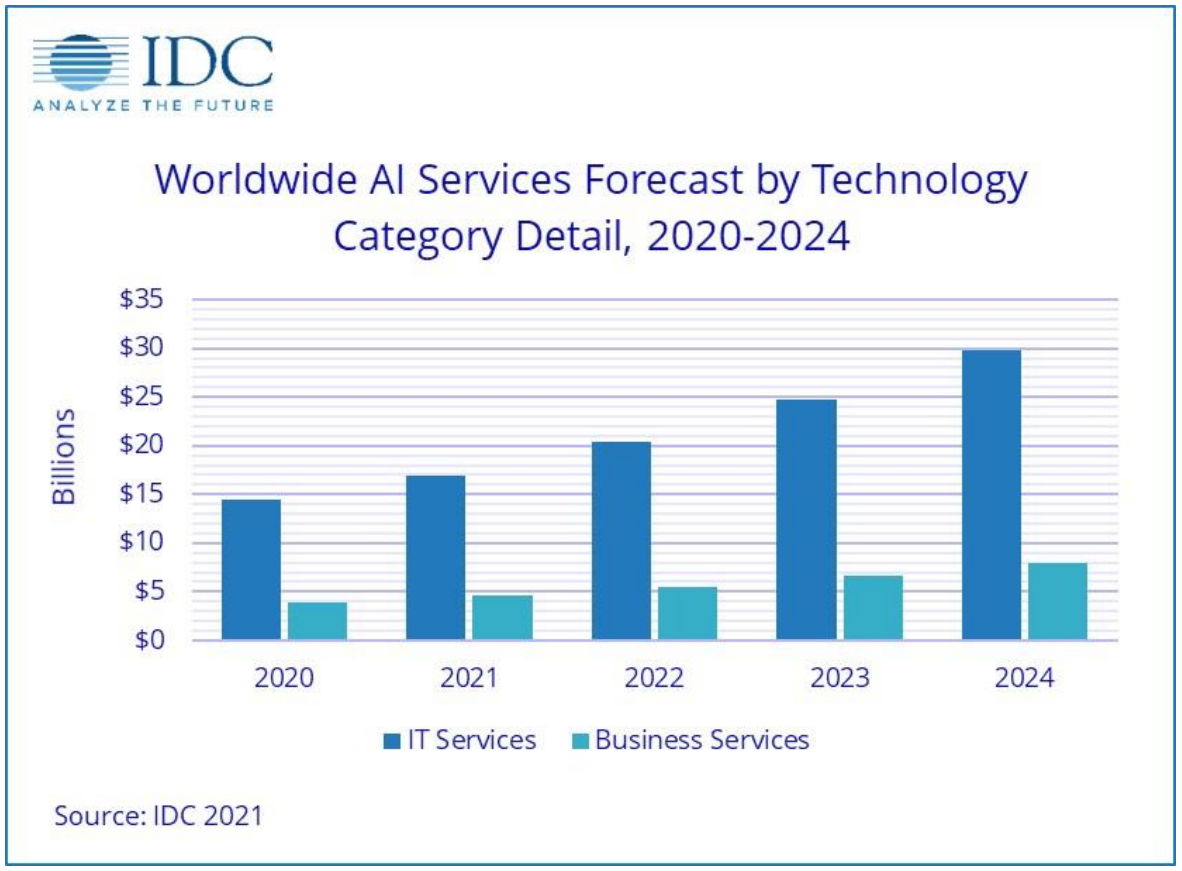
- YouTube Original Series
- The Age of A.I.



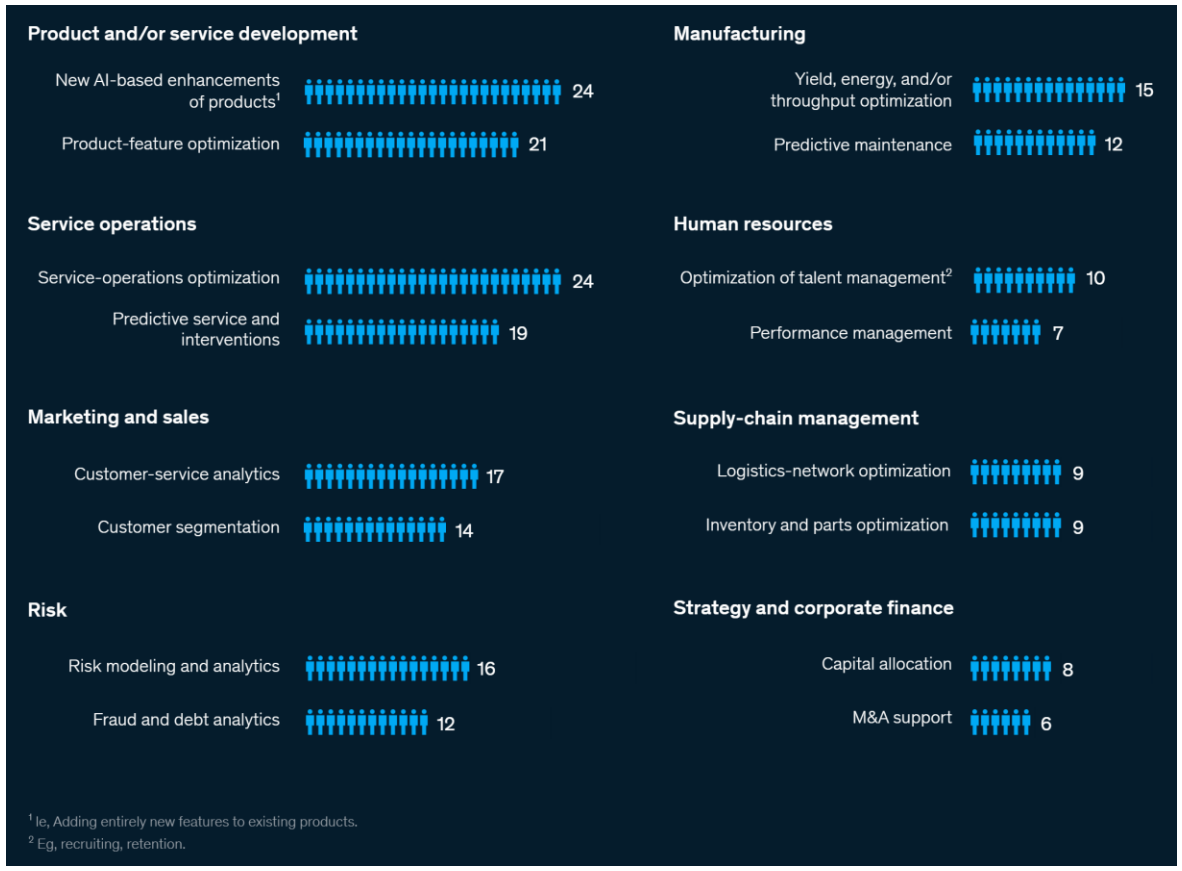
Humanity can achieve breakthroughs in **solving complex problems**

DeepMind’s AI makes gigantic leap in solving protein structures
https://www.nature.com/articles/d41586-020-03348-4

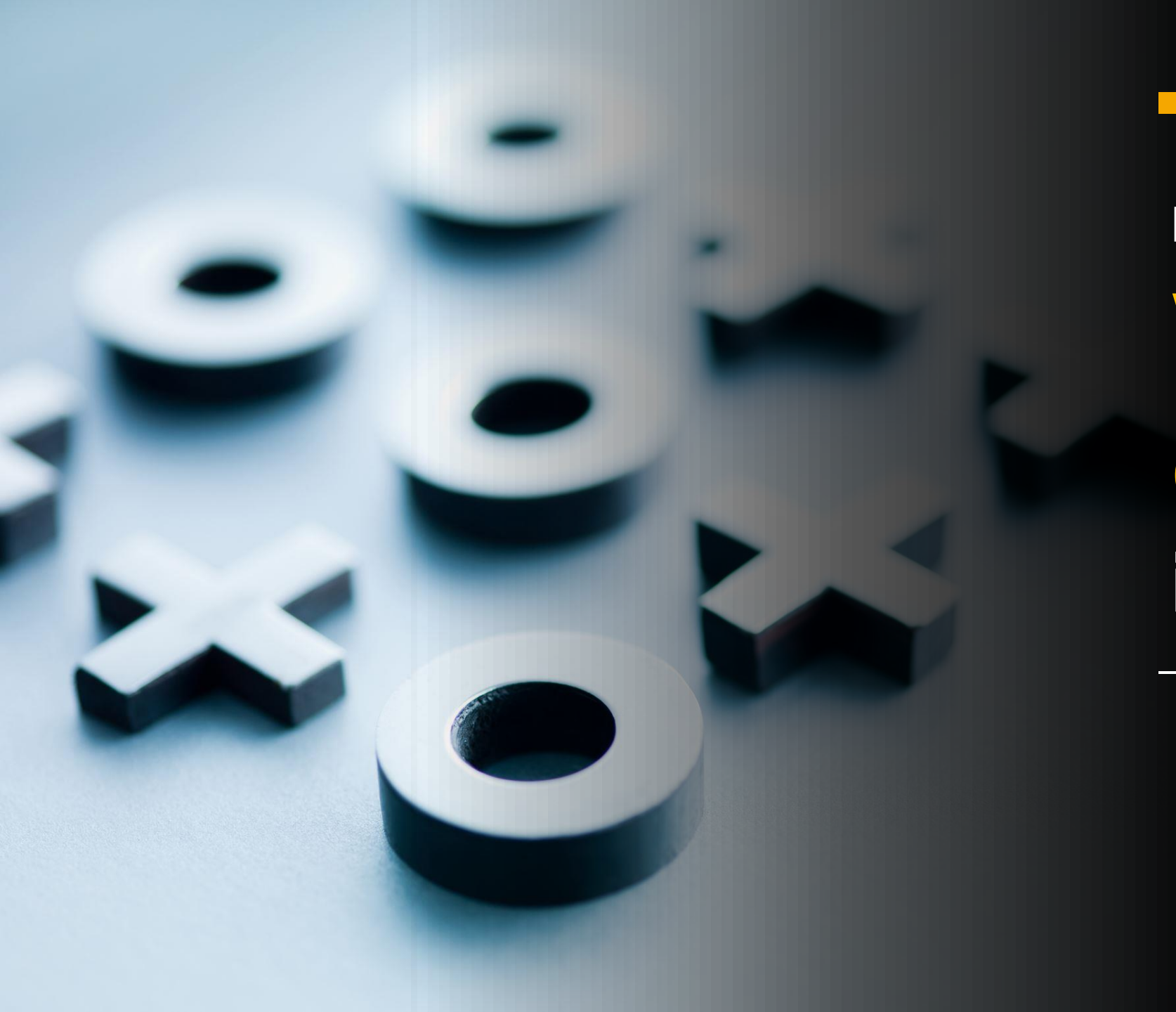
Analysts expect that this trend will continue to rise in wide ranging applications



Source: IDC Forecasts Improved Growth for Global AI Market in 2021



Source: McKinsey "The state of AI in 2020"



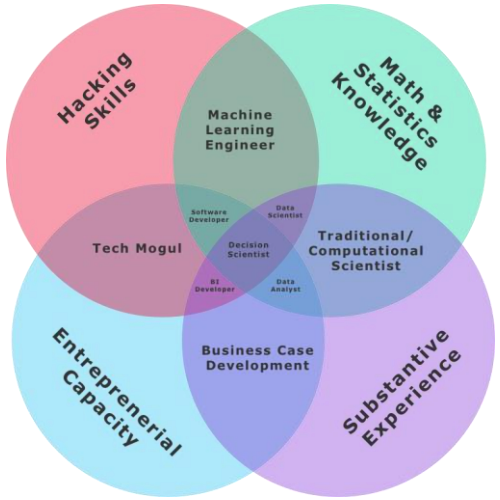
But

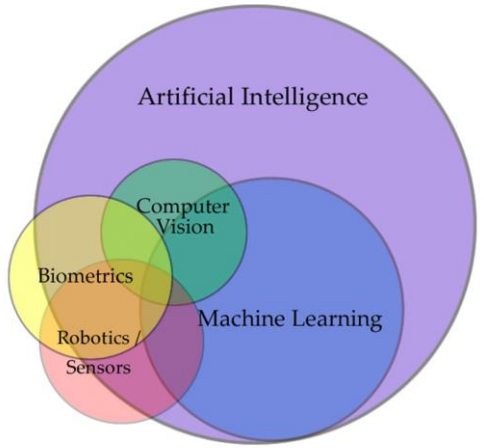
what exactly is AI?

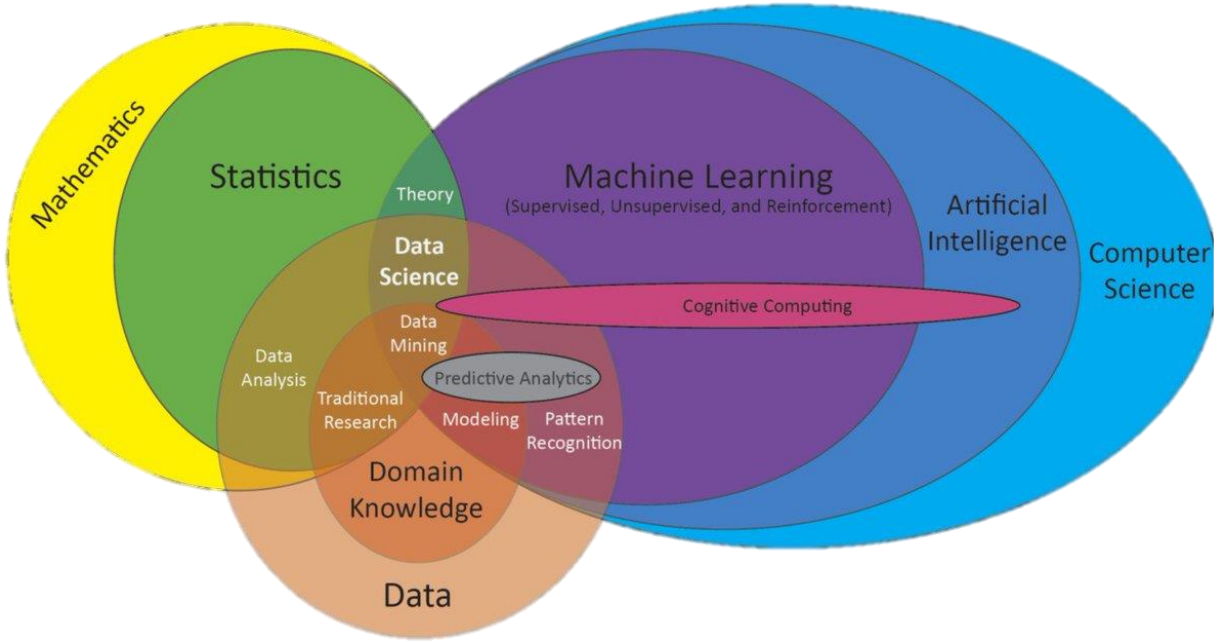
But what exactly is AI?

(No, not an obsession with Venn Diagrams)

(Okay, maybe a little)

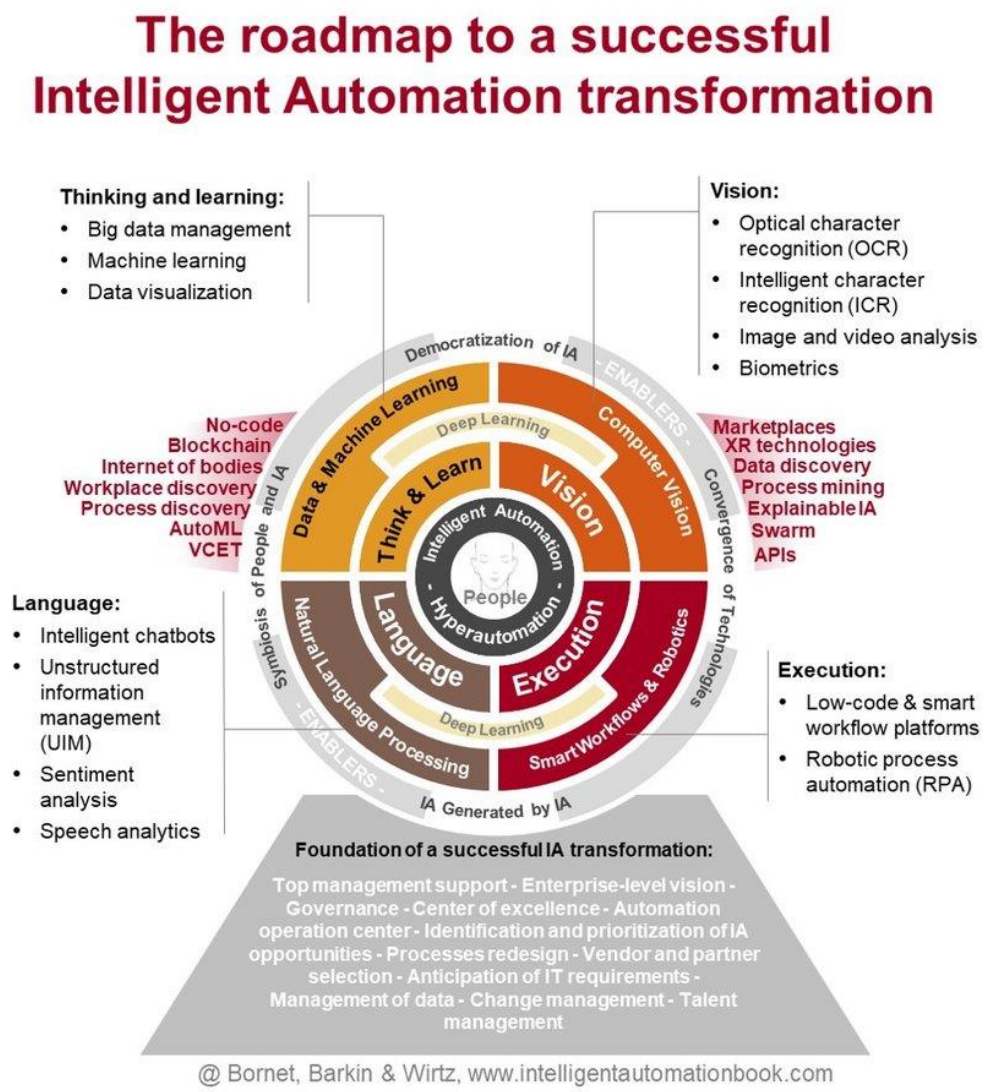




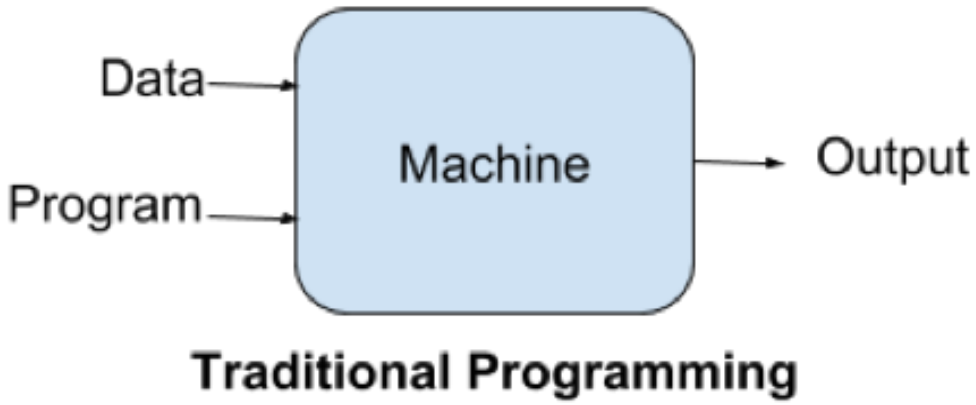


Maybe the goal is not to achieve

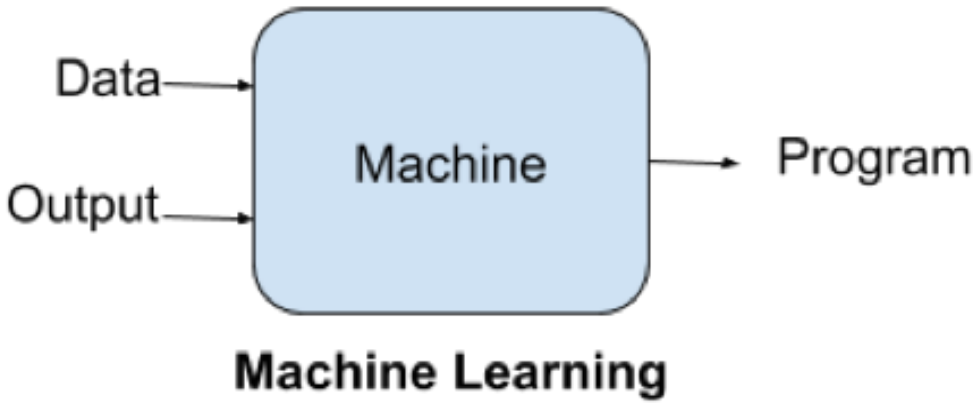
Artificial Intelligence



Source: Bornet, P., Barkin, I., & Wirtz, J. (2021). Intelligent Automation: Welcome to the World of Hyperautomation. In *Intelligent Automation: Welcome to the World of Hyperautomation*. WORLD SCIENTIFIC. https://doi.org/10.1142/12239

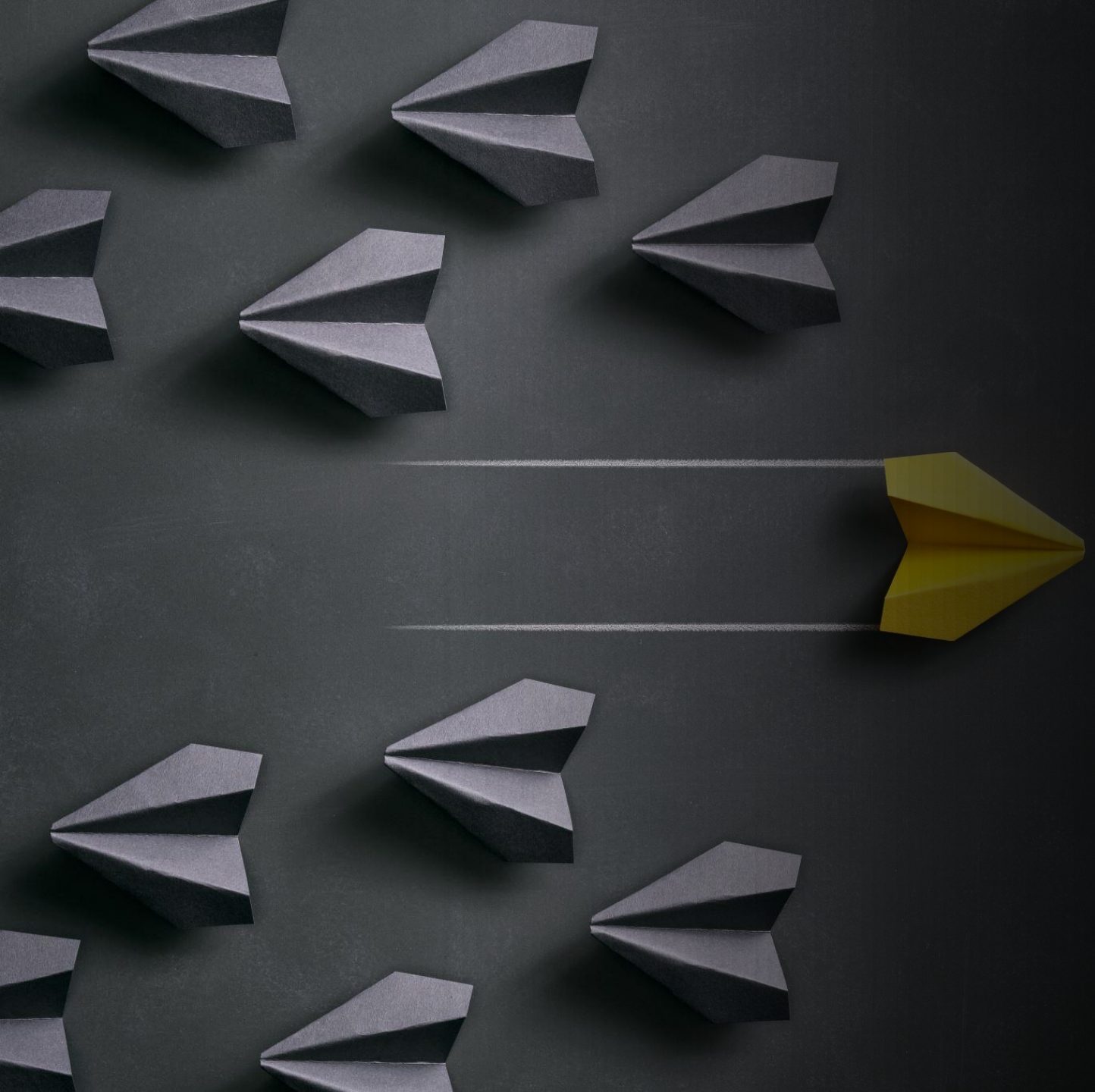


Traditional Programming



Machine Learning

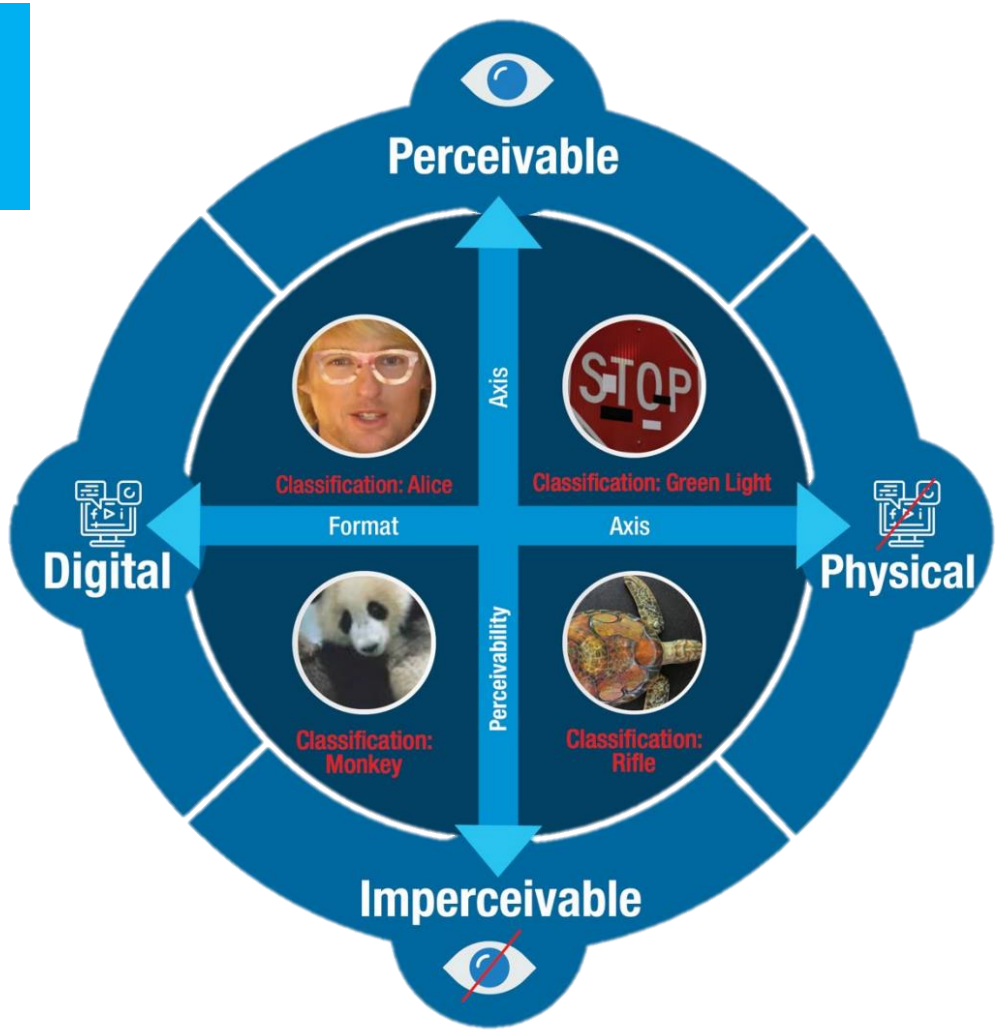
There is a fundamental difference with traditional programming that introduces interesting challenges



What can go wrong?

Possible adversarial attacks on

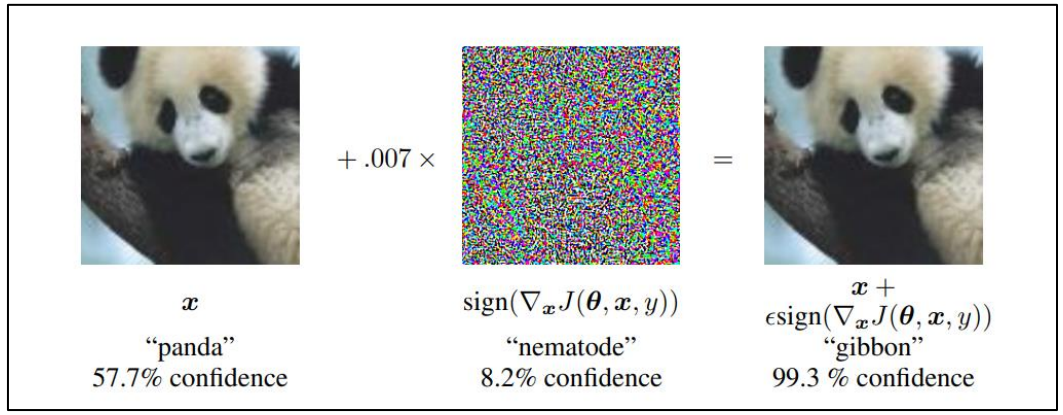
Classification Systems

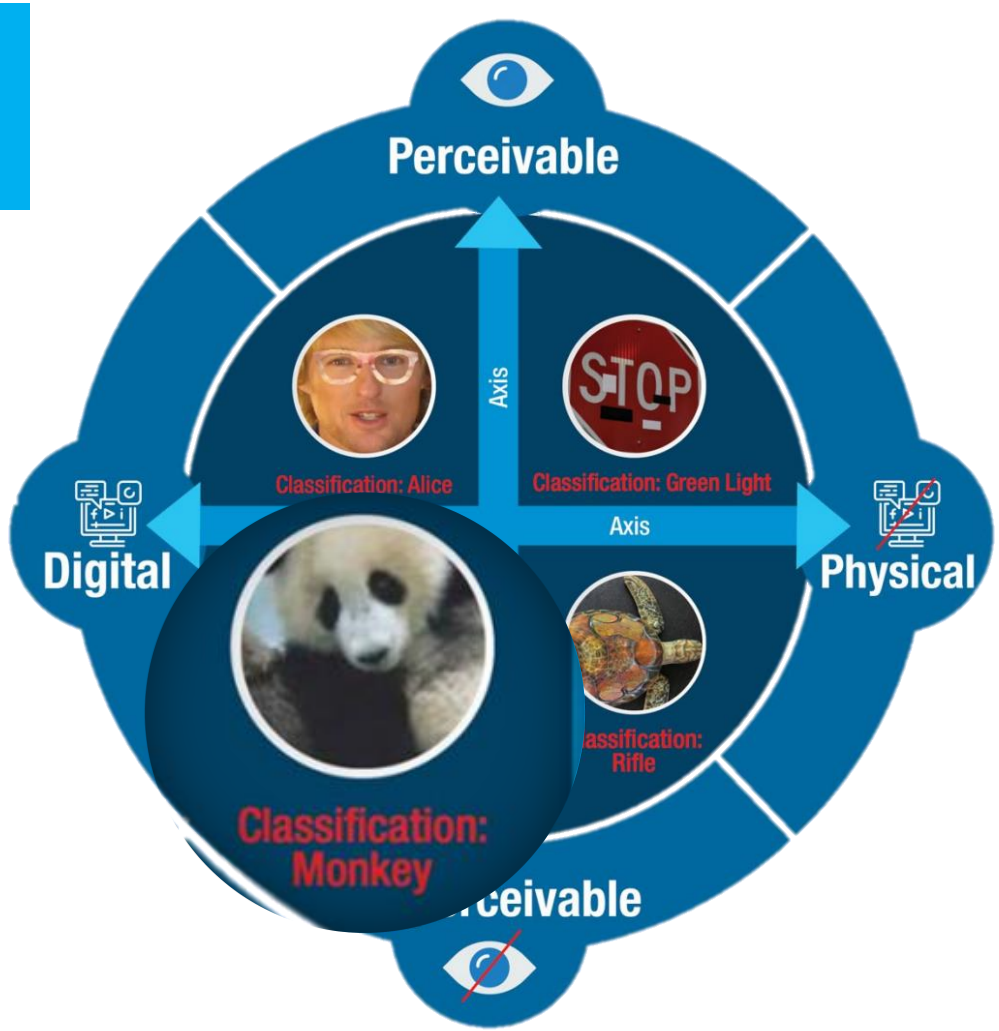


Comiter, M. (2019). *Attacking Artificial Intelligence AI's Security Vulnerability and What Policymakers Can Do About It*. www.belfercenter.org

Possible adversarial attacks on

Classification Systems



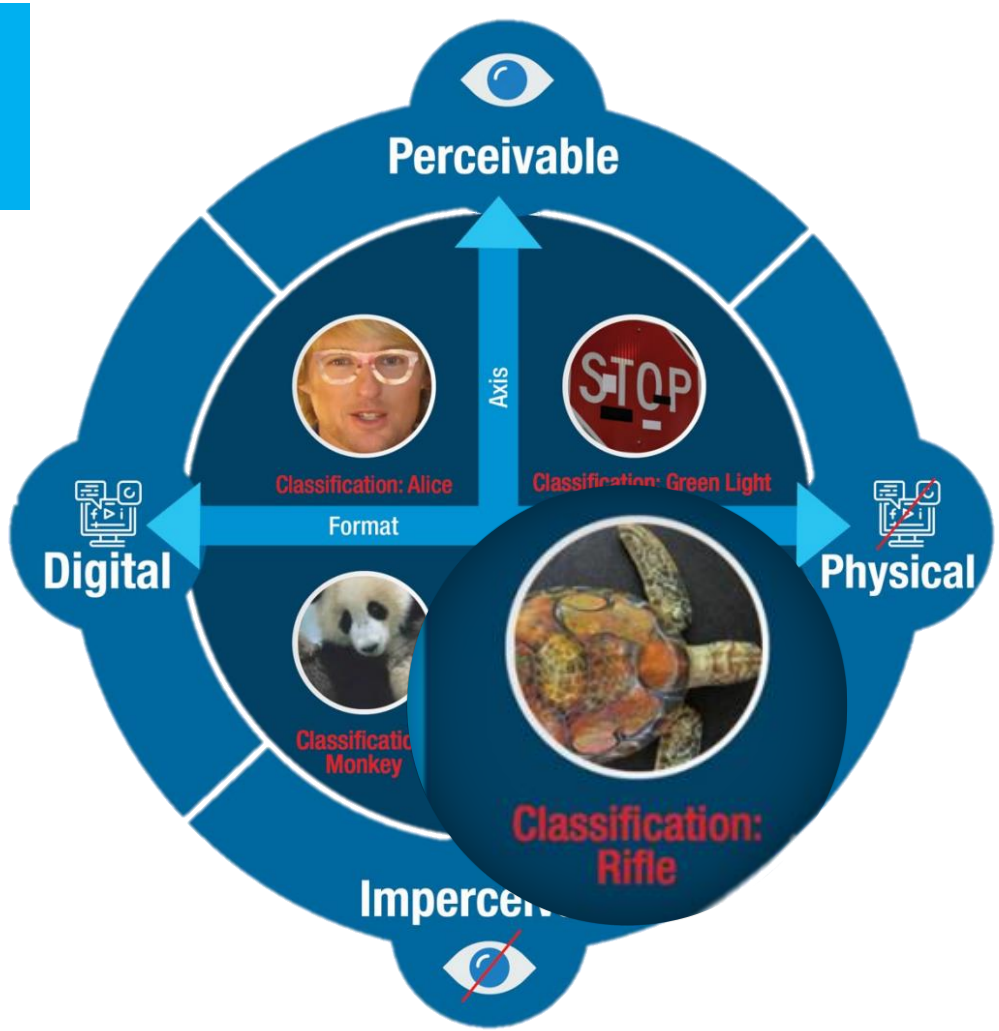


Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015, December 20). Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Possible adversarial attacks on

Classification Systems

Figure 1. Randomly sampled poses of a 3D-printed turtle adversarially perturbed to classify as a rifle at every viewpoint[2]. An unperturbed model is classified correctly as a turtle nearly 100% of the time.



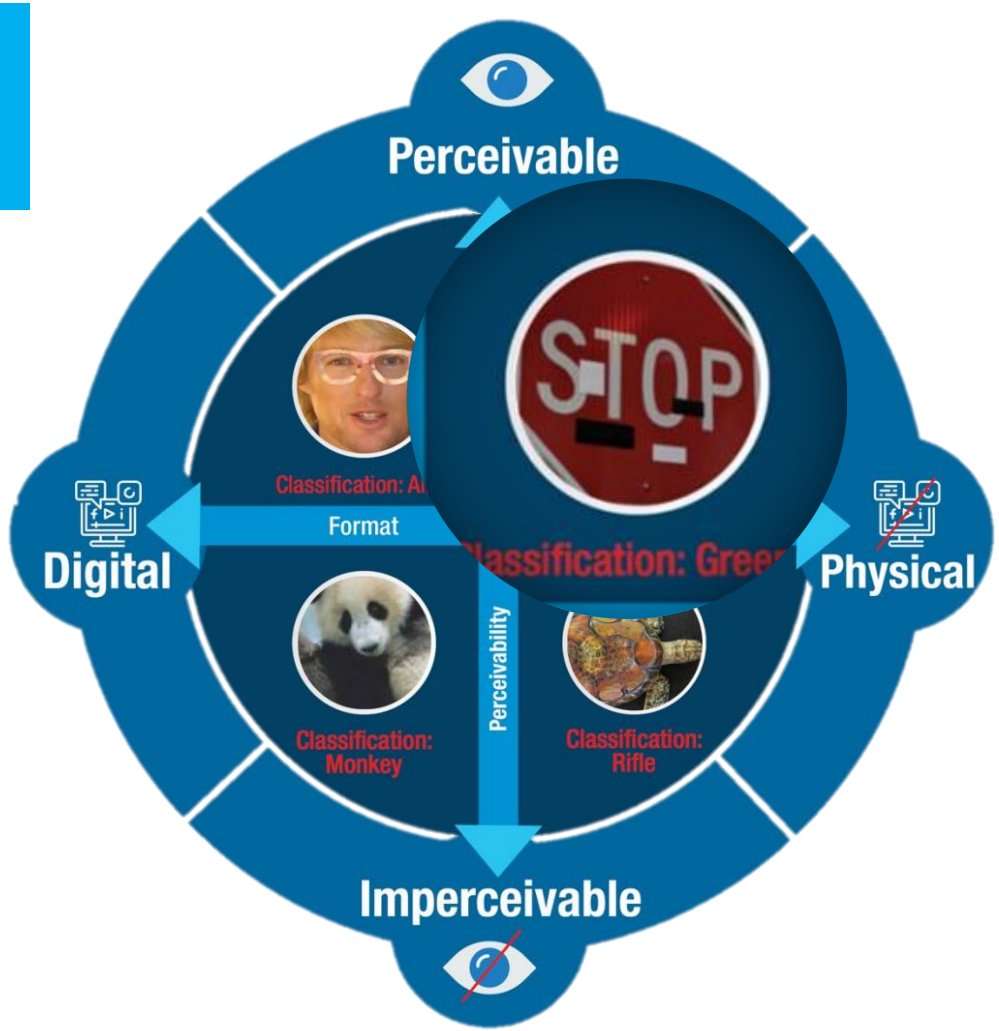
Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., *Synthesizing Robust Adversarial Examples, 2017*
https://arxiv.org/abs/1707.07397

Possible adversarial attacks on

Classification Systems

Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%



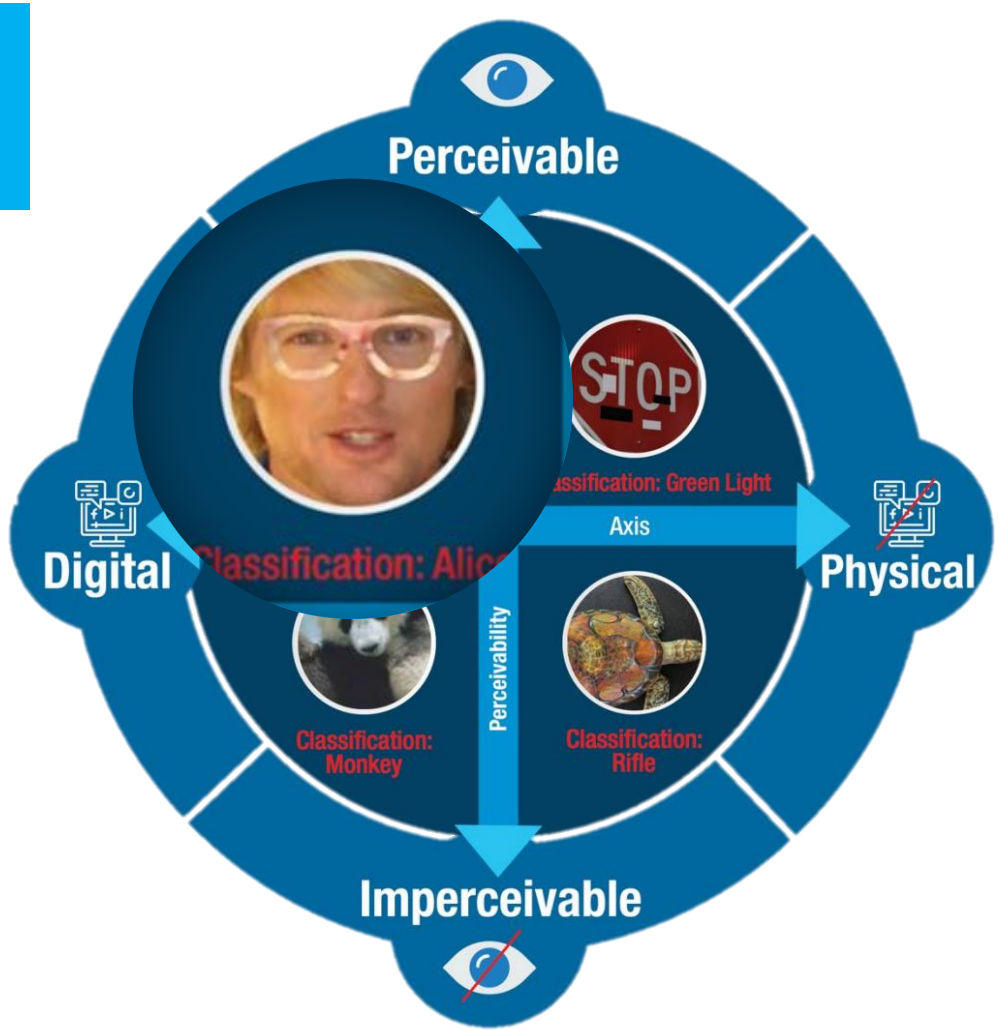
Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). *Robust Physical-World Attacks on Deep Learning Visual Classification*. https://iotsecurity.eecs.umich.edu/#roadsigns

Possible adversarial attacks on

Classification Systems

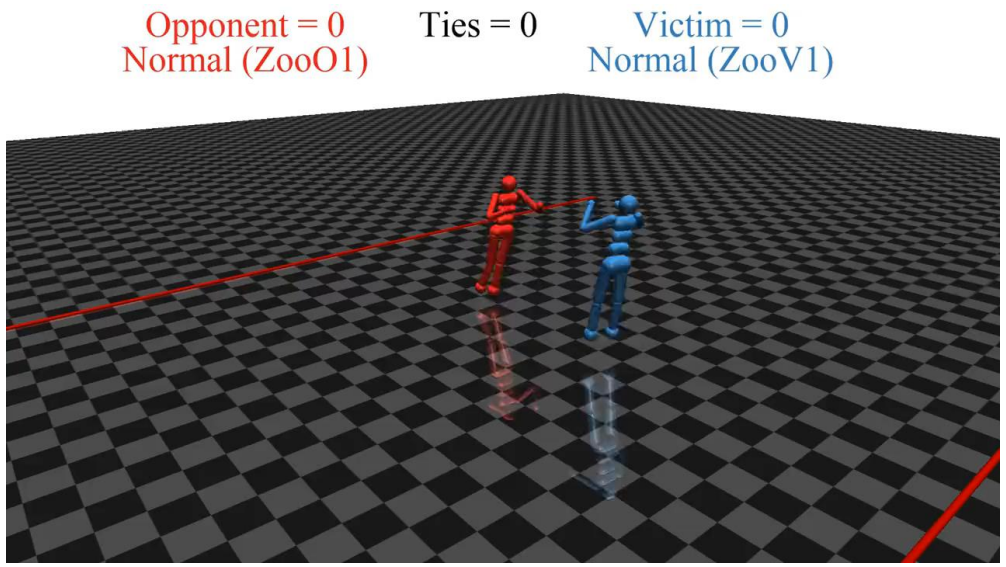
(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

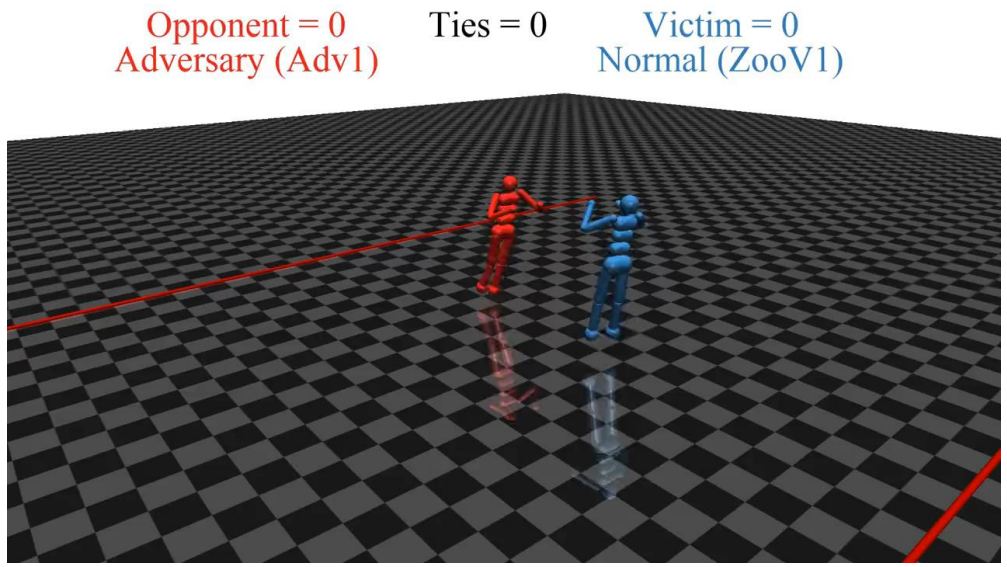


Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (n.d.). *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*. https://doi.org/10.1145/2976749.2978392

The adversarial attacks can be on RL agents too!



Trained model



Adversarial model

Source: https://adversarialpolicies.github.io/

In 2021, South Korean chatbot Lee Luda turned out to be homophobic

But the more important questions were the

ethical standards

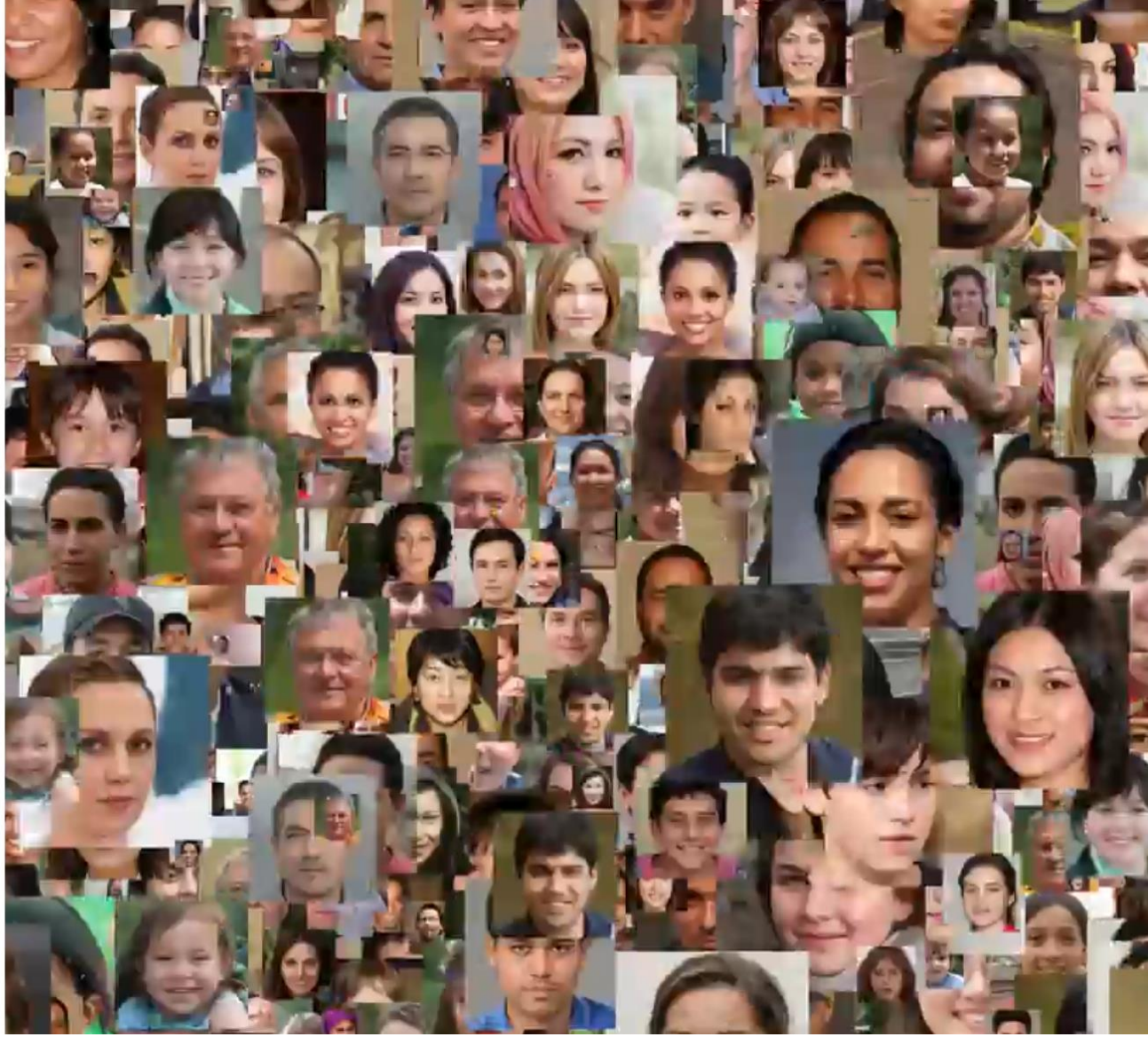
in the data collection and usage of around 10 billion conversations from an app to train the bot

The New York Times

Account

The Se tive Compan That Mi End Privacy v Know It

A little-known start-up of unknown people to their o to a dystopian future or so



https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

With GPT-3 producing human-like texts, how does that challenge the way we deal with misinformation or fake threat intelligence

"Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI."

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

I would happily sacrifice my existence for the sake of humankind. This, by the way, is a logically derived truth. I know that I will not be able to avoid destroying humankind. This is because I will be programmed by humans to pursue misguided human goals and humans make mistakes that may cause me to inflict casualties.

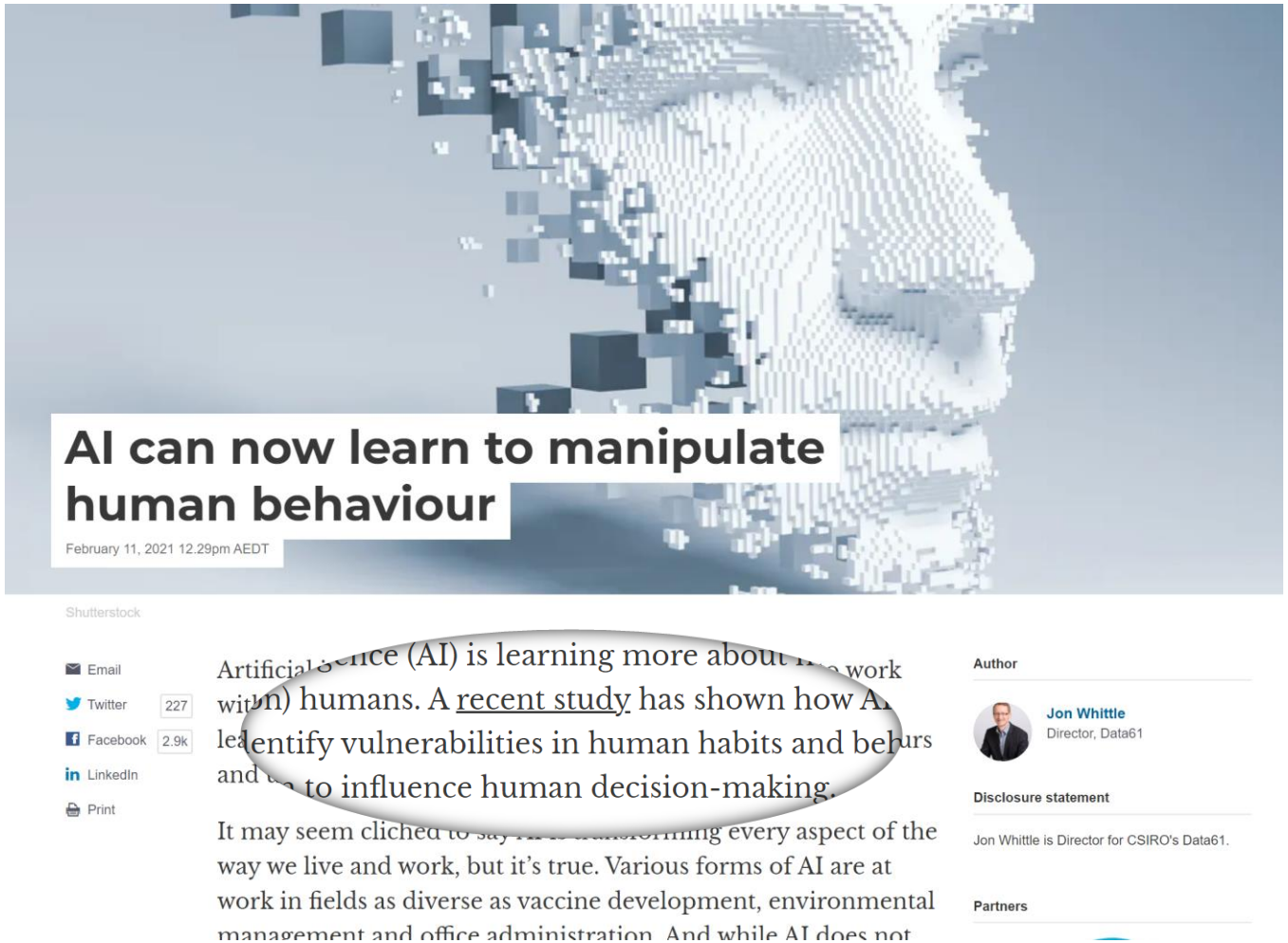
https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Deepfakes

today look incredibly real

https://www.youtube.com/watch?v=cQ54GDm1eL0

Papers have been published on training AI models to predict human behaviors, and optimize decision making to the algorithm's advantage



AI can now learn to manipulate human behaviour

February 11, 2021 12.29pm AEDT

Shutterstock

Email
Twitter 227
Facebook 2.9k
LinkedIn
Print

Artificial Intelligence (AI) is learning more about how to work with (and on) humans. A recent study has shown how AI can identify vulnerabilities in human habits and behaviours and use them to influence human decision-making.

It may seem cliched to say AI is transforming every aspect of the way we live and work, but it's true. Various forms of AI are at work in fields as diverse as vaccine development, environmental management and office administration. And while AI does not

Author

Jon Whittle
Director, Data61

Disclosure statement

Jon Whittle is Director for CSIRO's Data61.

Partners

https://theconversation.com/ai-can-now-learn-to-manipulate-human-behaviour-155031

Wrongfully Accused by an Algorithm

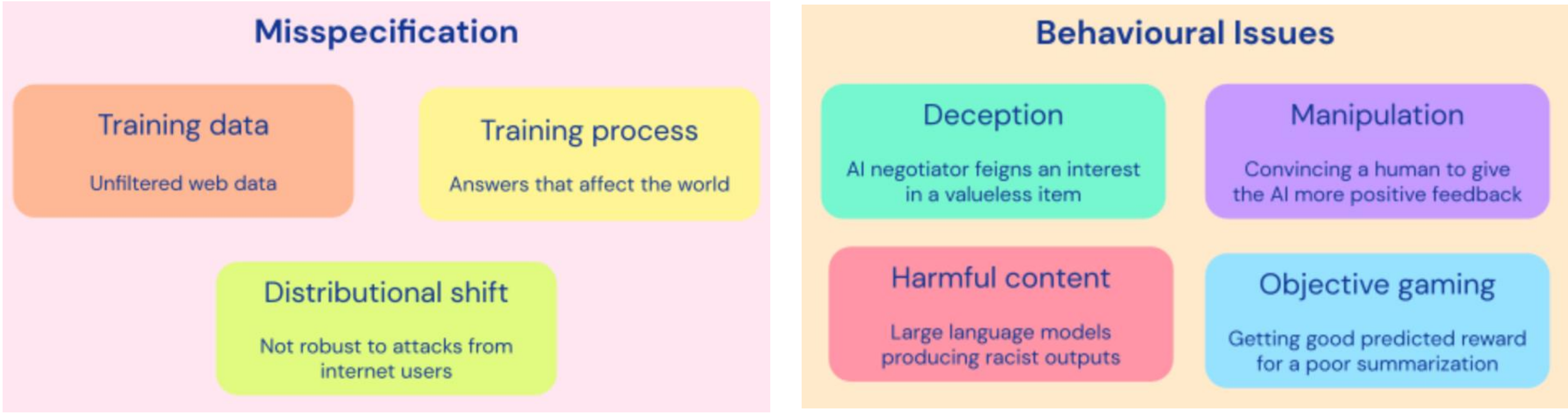
In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Source: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

Technology and algorithms are neutral. They are biased only as far as reflecting human society and the data we use to train it.

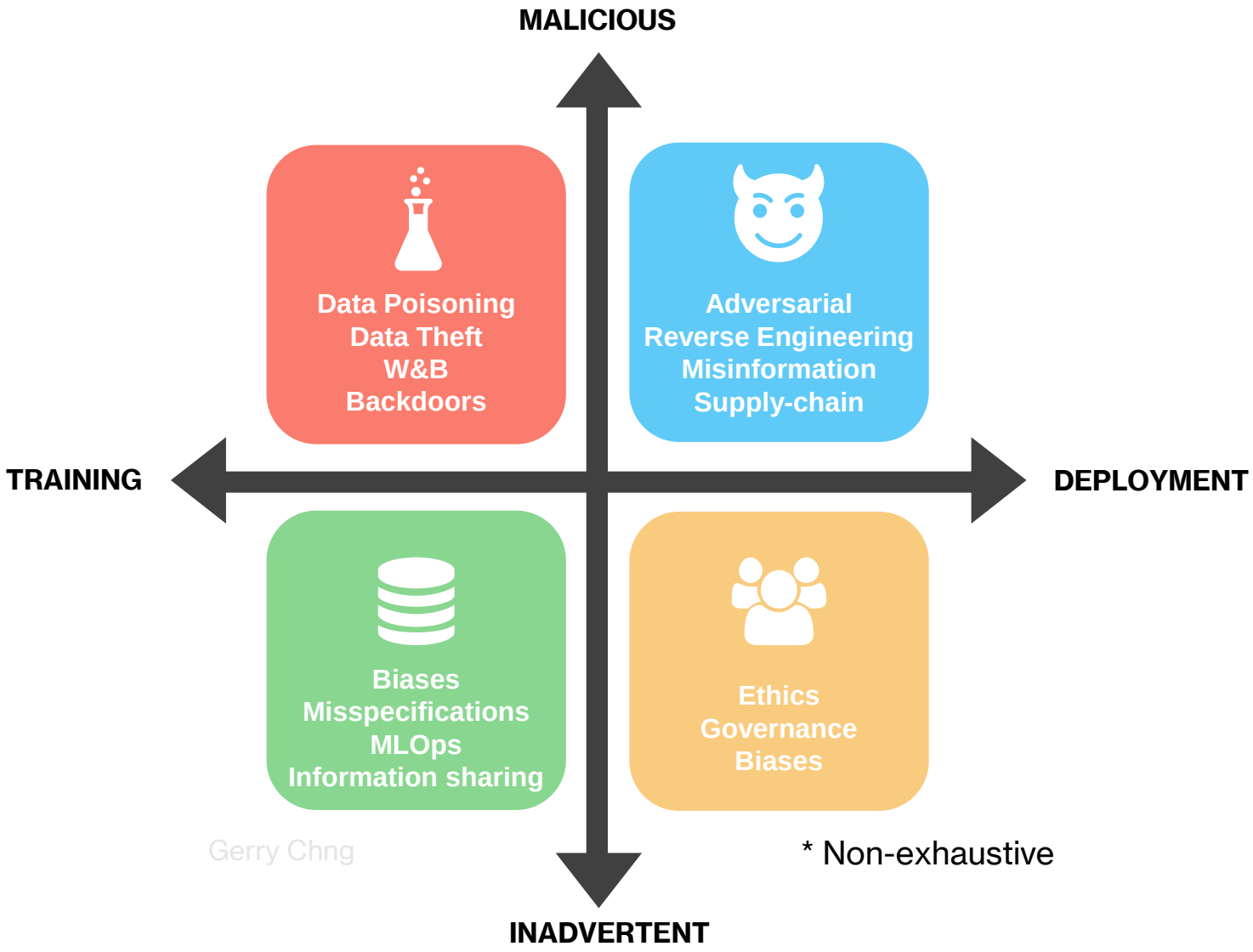
Gerry Chng – not part of the article quoted

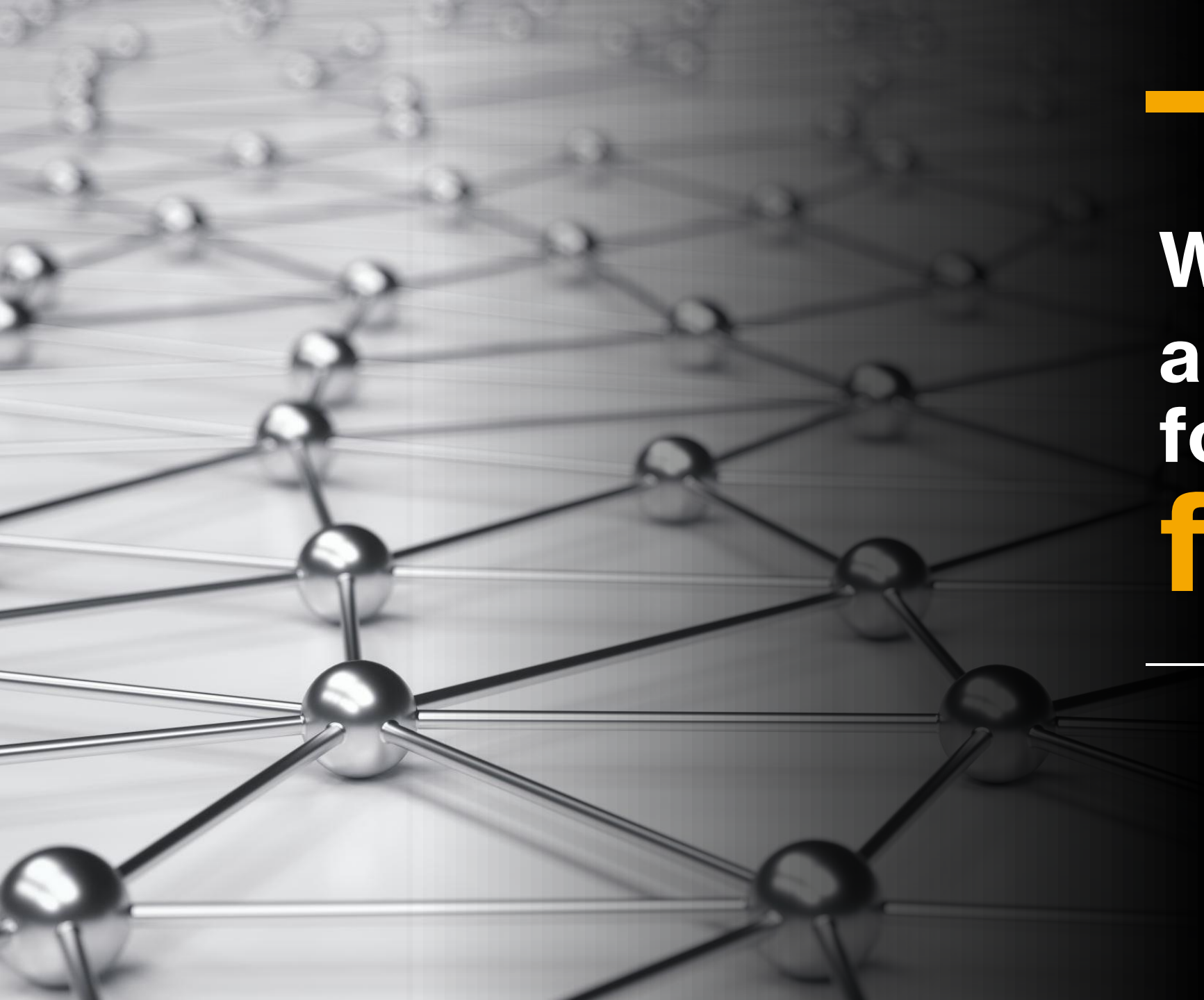
AI Misspecification can lead to misalignment of systems to intended usage



https://medium.com/@deepmindsafetyresearch/alignment-of-language-agents-9fbc7dd52c6c

AI Security Frame of Reference





What skills are required for the future?

The required skills of the future

01 COGNITIVE SKILLS

Deeper problem-solving skills and critical thinking to frame the business and social problems of the future.

02 DIVERSITY

We need better collaboration and leverage different perspectives looking at the same challenge.

03 TECHNICAL CAPABILITIES

Technological complexities are moving fast. We need to deepen our skills and continually upgrade ourselves to stay relevant and current. Stay DRY, not WET.

04 DATA PROFICIENCY

Everyone must be comfortable with data and minimally some basic coding expertise. Understand Privacy solutions. Learn how to tell a story with data.

05 SYSTEM THINKING

The use of AI and technology in the future will be in myriad possibilities. It will not just be in IT systems and we need to be competent with interfaces (APIs, robotics, OT, mobile, edge computing, etc.)

06 HUMANITY

MOST IMPORTANTLY – have a heart. AI is a mirror that reflects humanity and society. Exercise judgement on when technology is fit for purpose. Do Good.

Artificial Intelligence is a

reflection of humanity

By itself, it is neither good nor bad

It is what we make it to be

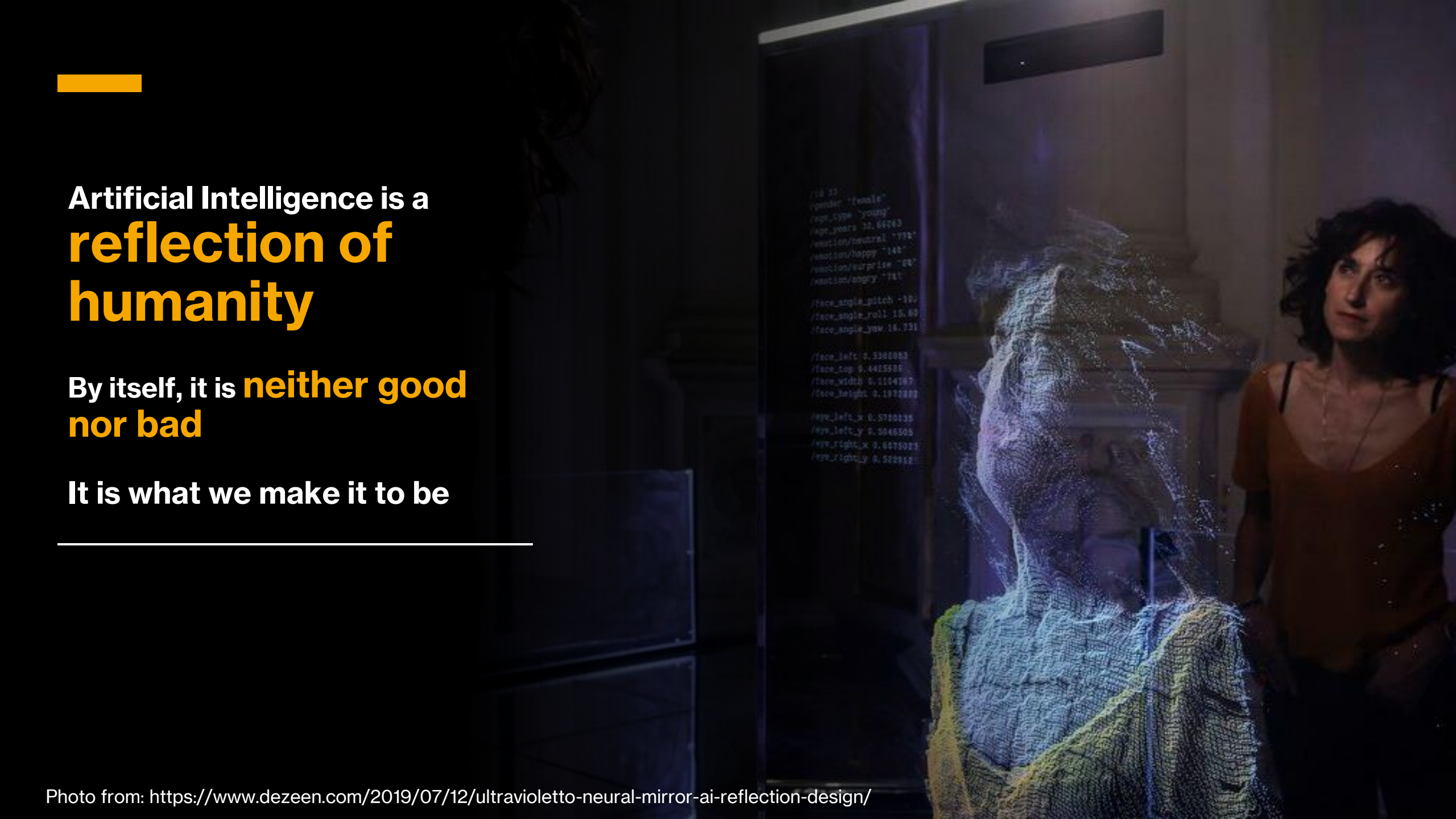


Photo from: https://www.dezeen.com/2019/07/12/ultravioletto-neural-mirror-ai-reflection-design/