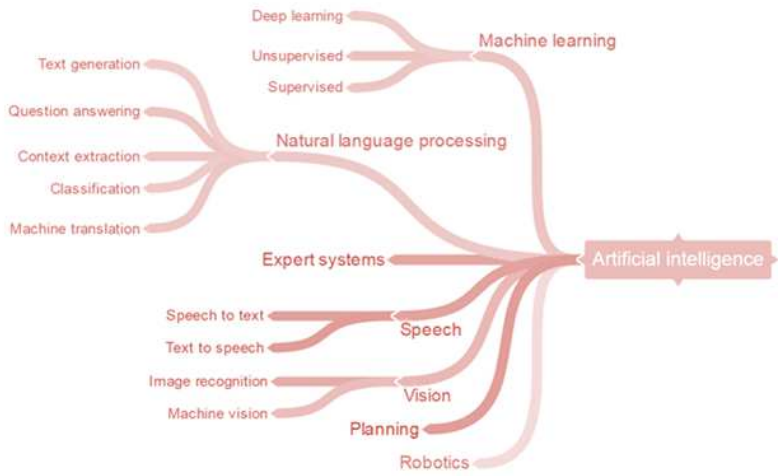
Black-box algorithms: Blind trust?

Mona de Boer

pwc



AI that can sense...

Hear | See | Speak | Feel

- Natural language
- Audio & speech
- Machine vision
- Navigation
- Visualisation

AI that can think...

Understand | Assist | Perceive | Plan

- Knowledge & representation
- Planning & scheduling
- Reasoning
- Machine learning
- Deep learning

AI that can act...

Physical | Creative | Cognitive | Reactive

- Robotic process automation
- Deep question & answering
- Machine translation
- Collaborative systems
- Adaptive systems

Statistics | **Econometrics** | **Optimisation** | **Complexity theory** | **Computer science** | **Game theory**

FOUNDATION LAYER

AI ranges from hardwired automation to fully autonomous intelligence

Human in the loop

No human in the loop

Hardwired systems

Assisted Intelligence

Using data and analytics to drive business insights within existing decisions and actions

Automation

Automating business processes previously performed by humans

Adaptive systems

Augmented Intelligence

New ways for computers and humans to collaborate in making better decisions and taking more effective actions

Autonomous Intelligence

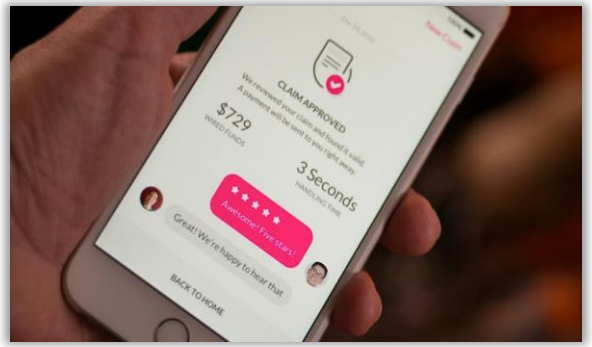
Systems that are adaptive and can autonomously carry out tasks without human intervention

Source: *Inside Sundar Pichai's Plan to Put AI Everywhere*, Forbes May 2016

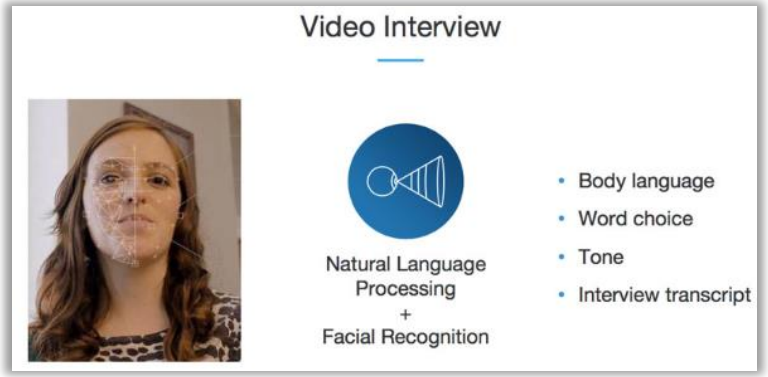
AI is disrupting the entire value chain by automating existing processes, uncovering new value from data and augmenting human decisions and actions

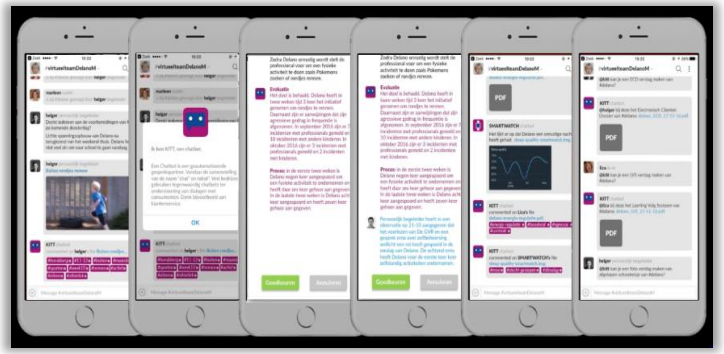
‘Marketable’ insights

Lowering cost

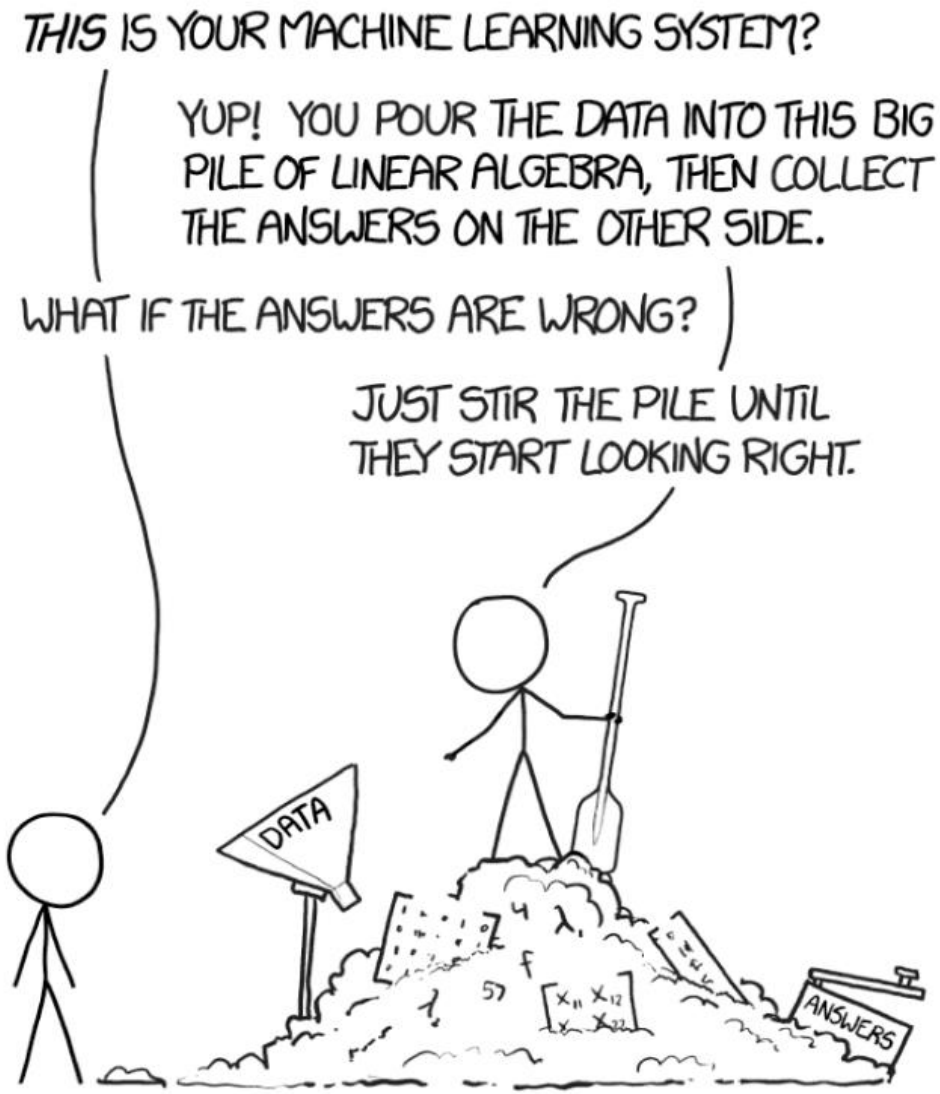


Improved decisions

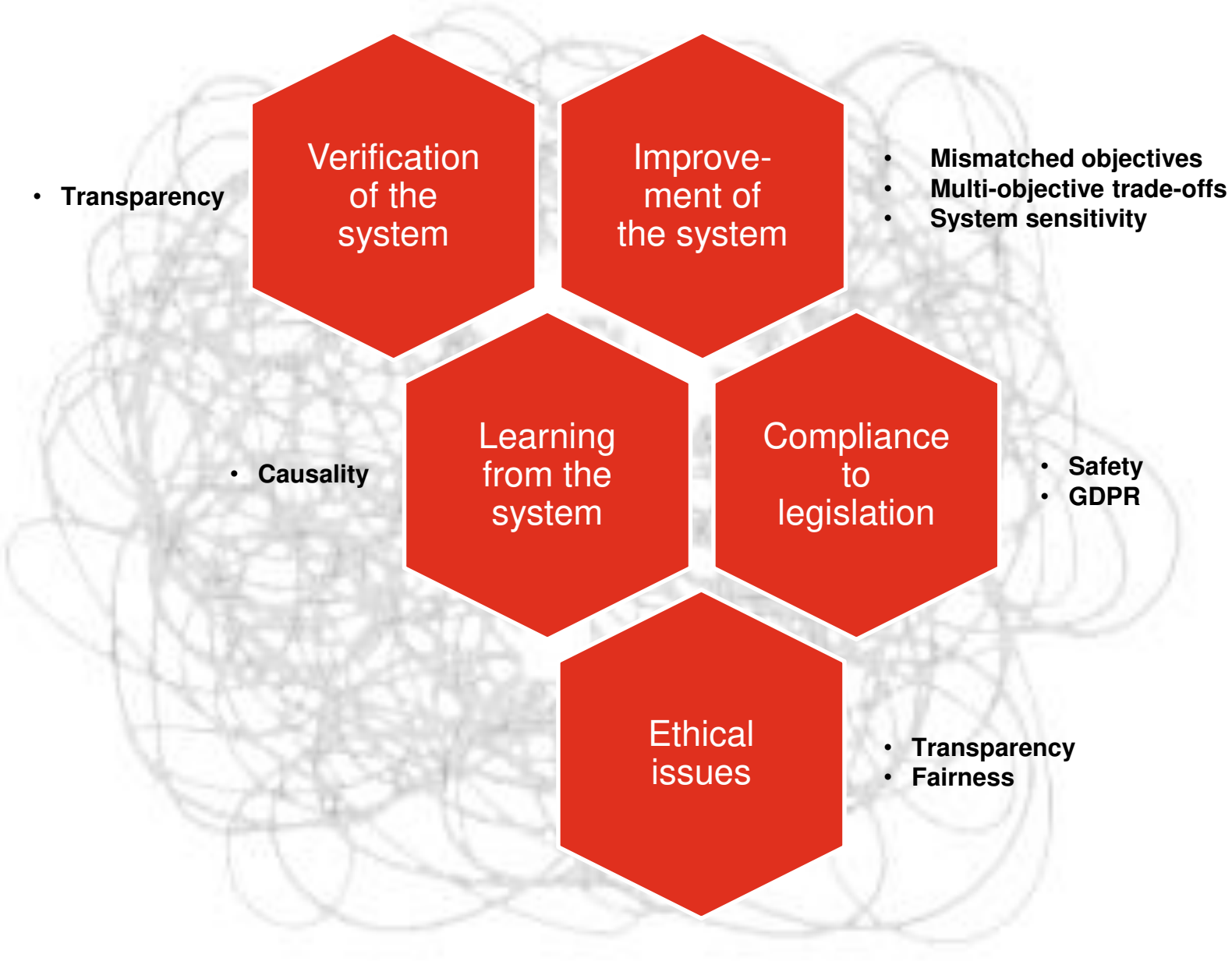




Let’s presume we’ll become consciously competent at ‘doing AI’…



… then we'll need XAI for…



Source: *Su et al, 2018*

… as a basis to trust AI

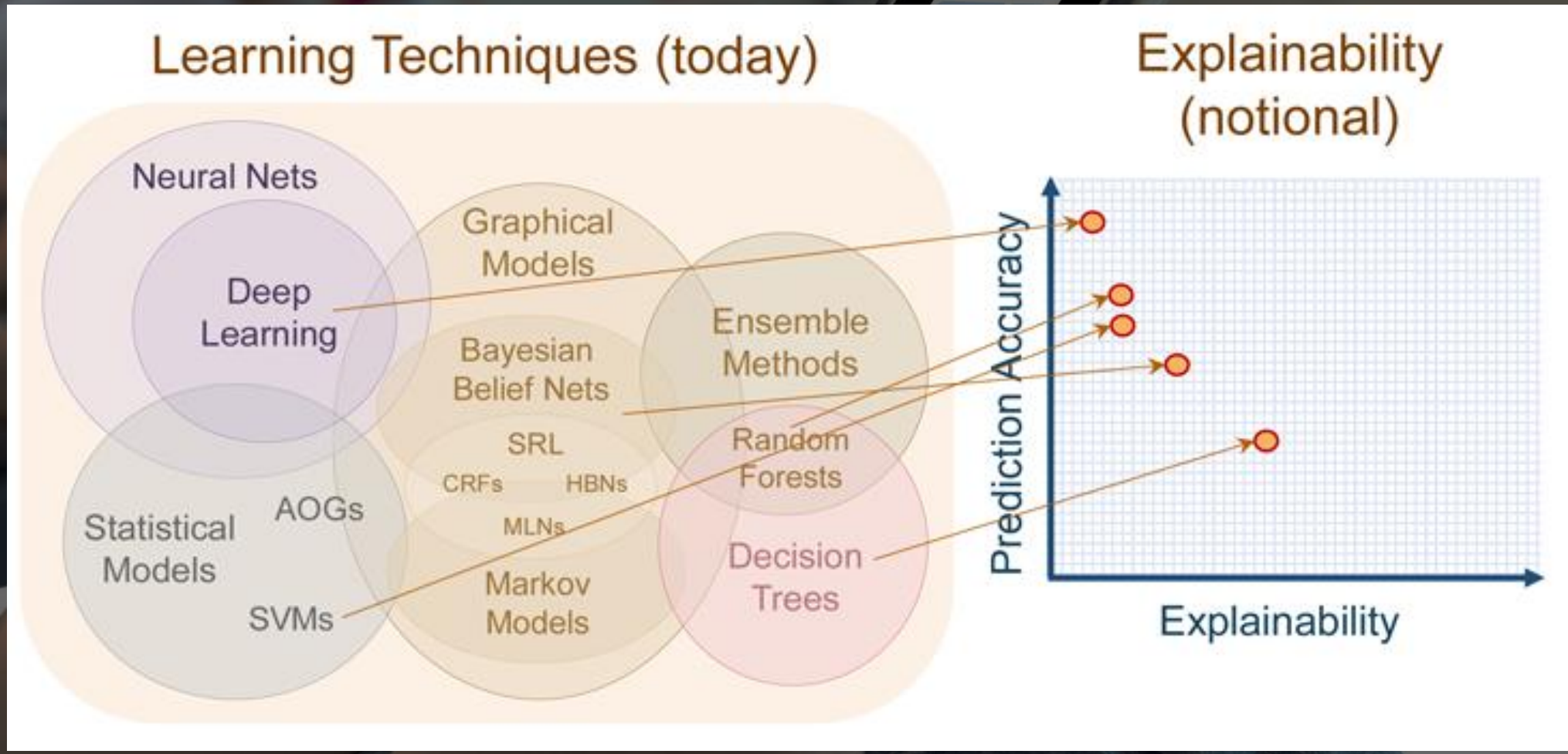
… call for action today

Despite finding themselves in uncharted territory, executives surveyed said trust tops the AI agenda. And they're taking steps to manage that, including using AI itself to address risks like cyberthreats.

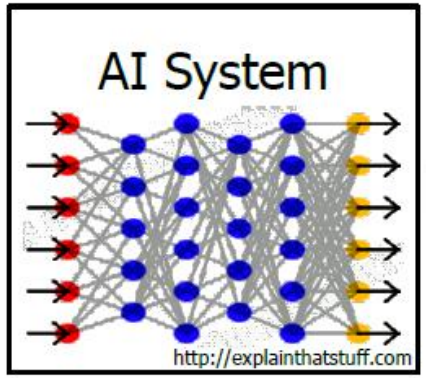
37% of executives said ensuring AI systems were trustworthy was their top priority

	All	Consumer Markets	Energy, Utilities, Mining	Financial Services	Health	Industrial Products	Tech, Media, Telecom
Boost AI security with validation, monitoring, verification	64%	70%	72%	60%	58%	63%	70%
Create **transparent, explainable, provable** AI models	61%	58%	67%	64%	56%	60%	63%
Create **ethical, legal, understandable** AI systems	55%	52%	57%	53%	56%	53%	58%
Improve governance with AI operating models, processes	52%	55%	63%	54%	36%	51%	60%
Test for bias in data, models, human use of algorithms	47%	50%	43%	48%	48%	48%	50%
Use **AI to manage risk, fraud, cybersecurity** threats	46%	47%	41%	52%	55%	34%	49%

Source: PwC 2019 AI Predictions
Bases: Total, 1,001; Consumer Markets, 132; Energy/Utilities/Mining, 46; Financial Services, 187; Health, 110; Industrial Products, 208; Tech/ Media/Telecom, 208; Other, 110.
Qs: What steps will your organization take in 2019 to develop and deploy AI systems that are trustworthy, fair, and stable? Which AI applications will be most important to your organization in 2019?



Source: *DARPA, 2016*



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

Transportation

Finance

Security

Legal

Medicine

Military

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Source: *DARPA, 2016*

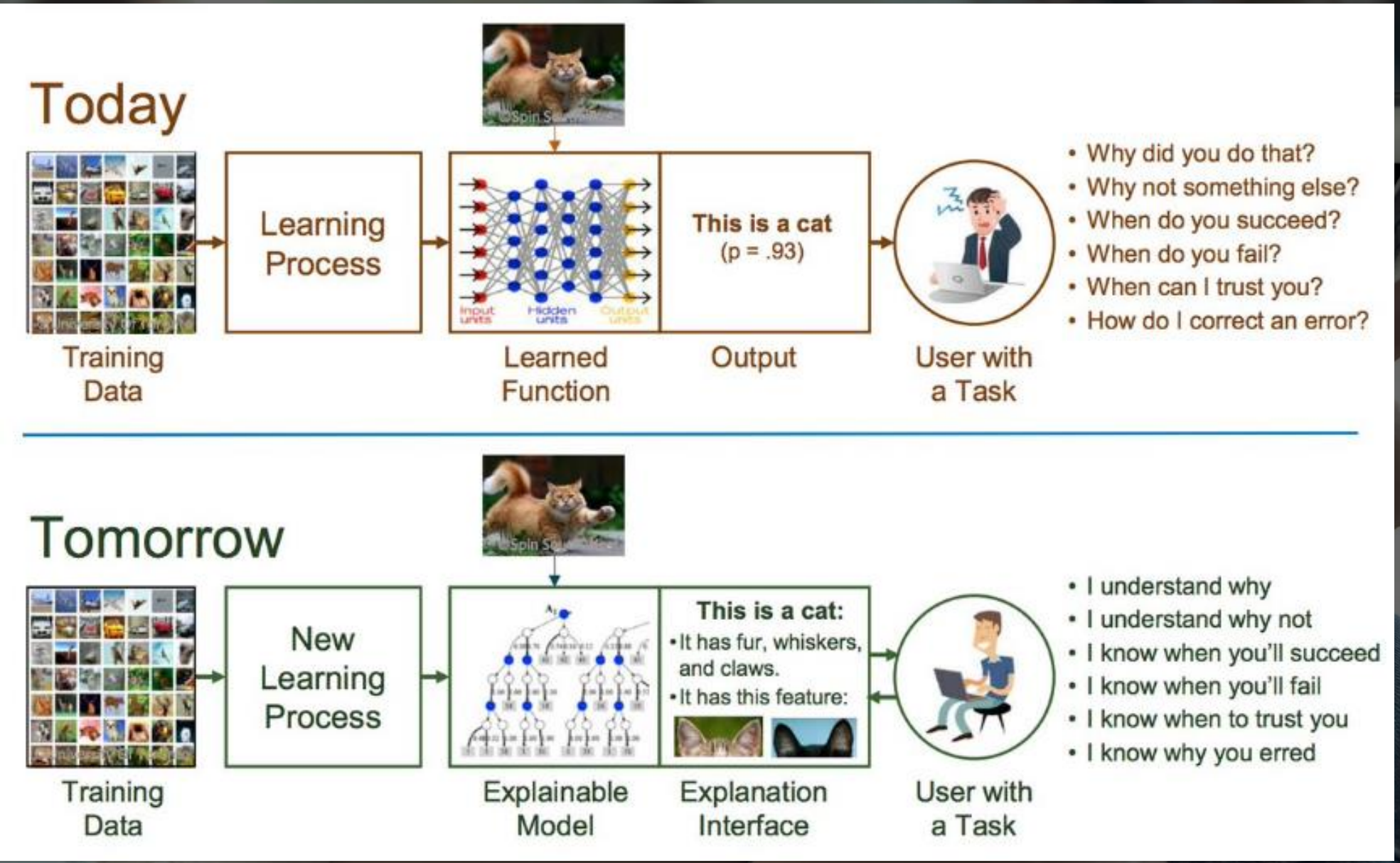
GDPR Art. 22 – Automated individual decision-making, including profiling

1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2) Paragraph 1 shall not apply if the decision:
   a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
   c) is based on the data subject's explicit consent.

3) In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

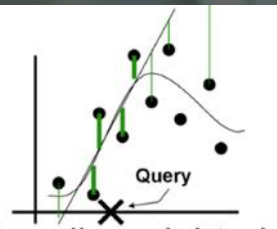
4) Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.



Source: *DARPA, 2016*

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52



Locally weighted regression

Explanation

P() = 0.54

P() = 0.07

P() = 0.05

Source: *Ribeiro et al, 2016*

Q: Is this a healthy meal? Textual Justification Visual Pointing

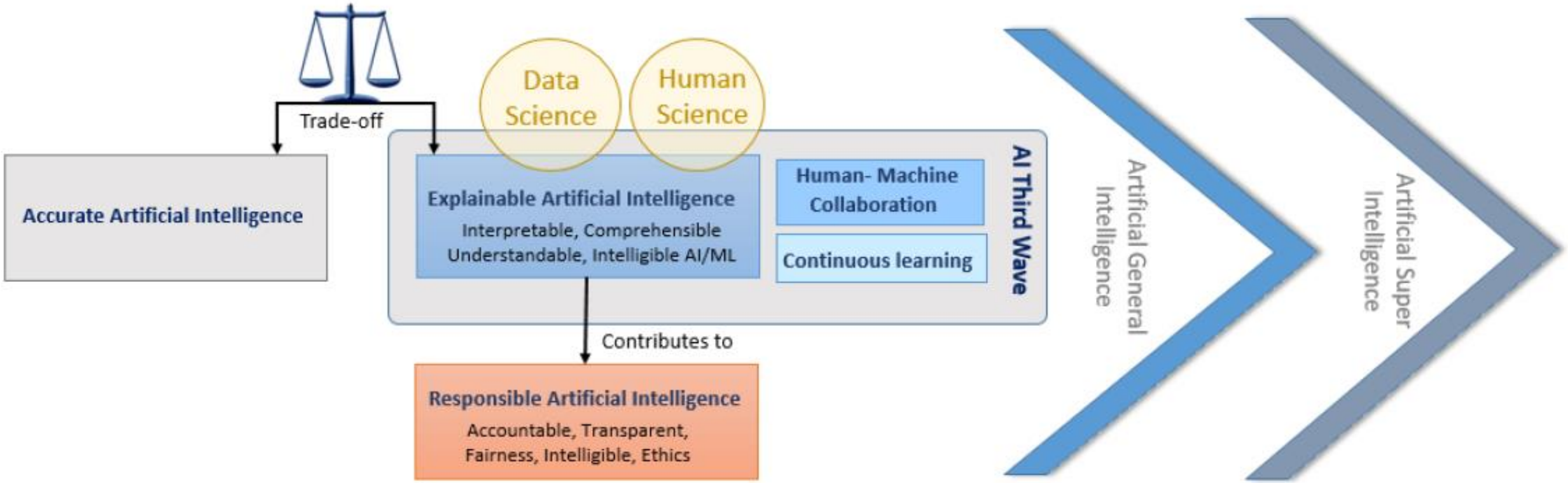
A: No

...because it is a hot dog with a lot of toppings.

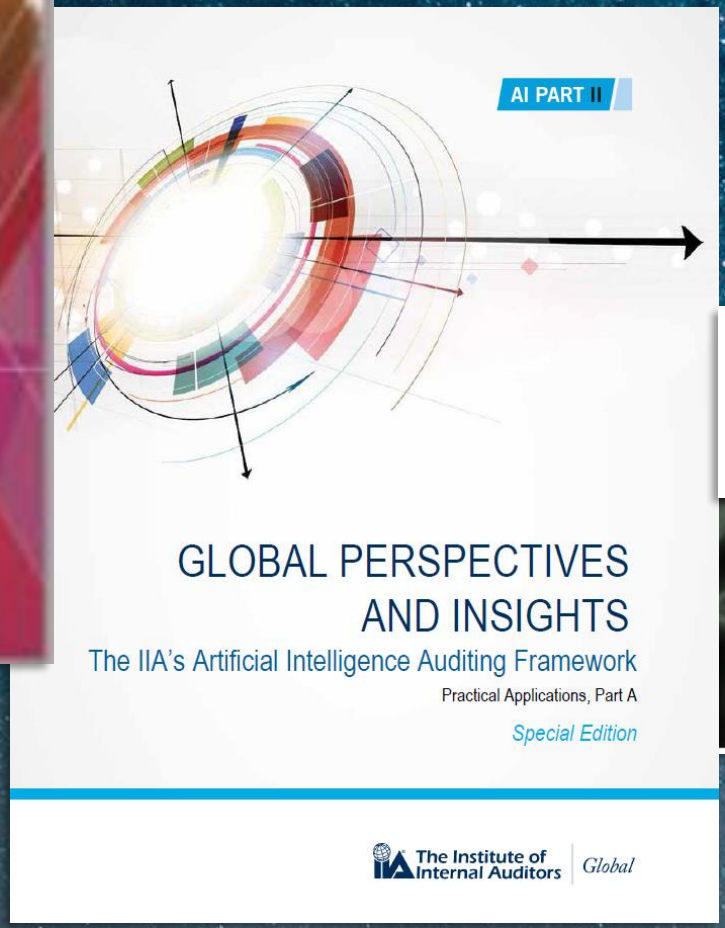
A: Yes

...because it contains a variety of vegetables on the table.

Source: *Park et al, 2018*



Source: *Adadi & Berrada, 2018*



NOREA
DE BEROEPSORGANISATIE VAN IT-AUDITORS

Algorithm Assurance – Nieuwe werkgroep ingesteld

14 december 2018