UNIVERSITY of BRADFORD

School of Computer Science, AI & Electronics

BCS West-Yorkshire Branch Webinar

bcs The Chartered Institute for IT

Fundamental Concepts in Artificial Intelligence for Real World Applications

Review of Responsible AI: key concepts, risks, and opportunities in the digital economy and Generative AI, with a focus on sustainability

This talk is inspired by the philosophical discussions with my wife Iustina Neagu

Professor Ciprian Daniel Neagu

Professor of Computing
AI Research (AIRE) Group Leader
Turing University Network institutional academic liaison
School of Computer Science, AI & Electronics
University of Bradford

Outline

Topic: context and challenges around using AI with the potential concerns and opportunities

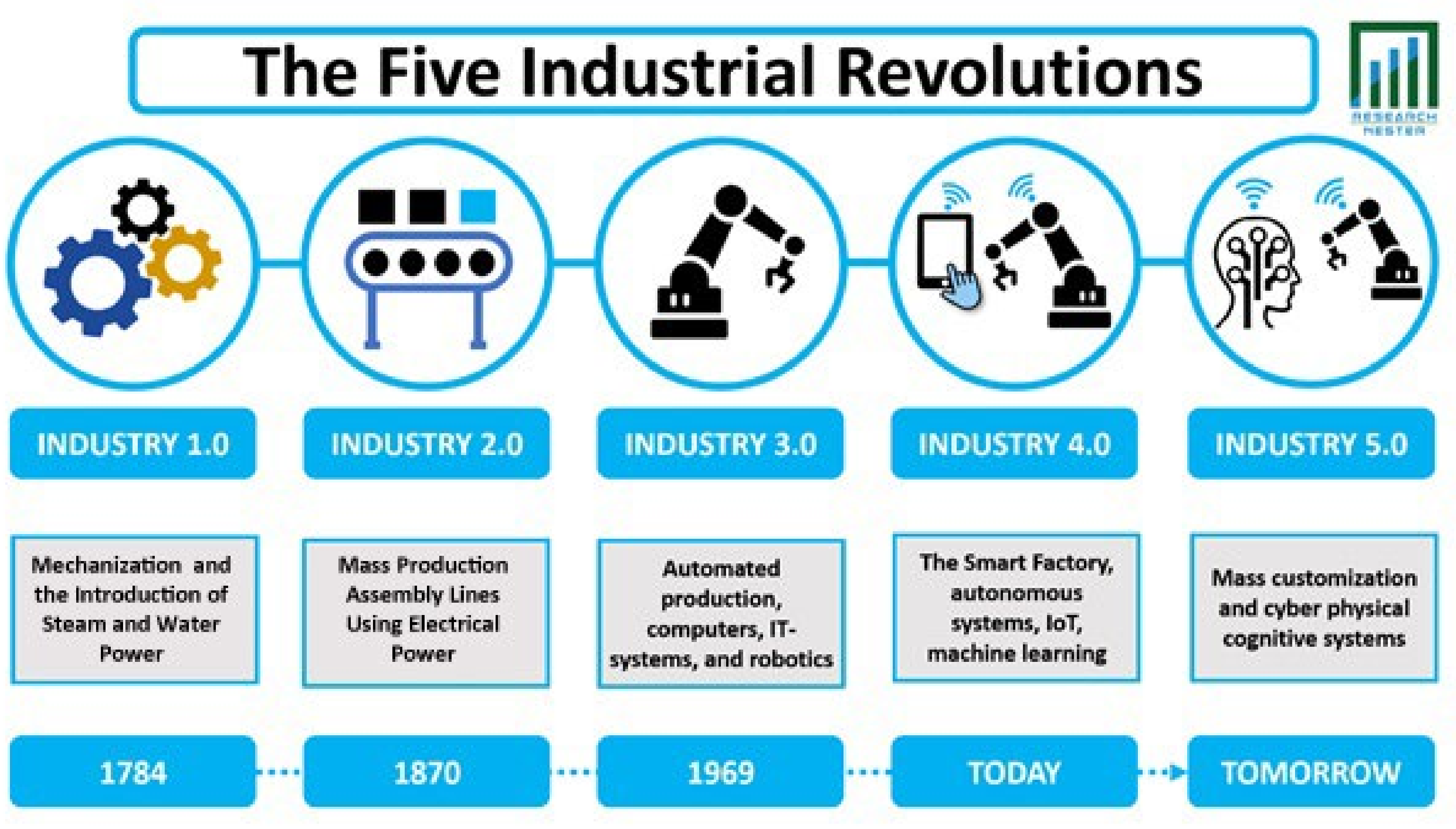
In this *Friday evening* presentation we will:

- Introduce:
  - The context and main fundamental concepts that are foundations of Ethics wrt Responsible AI
- Review:
  - State of the Art and Context
  - Expectations and Opportunities
  - Challenges: access rights and insights using AI
  - Dilemmas

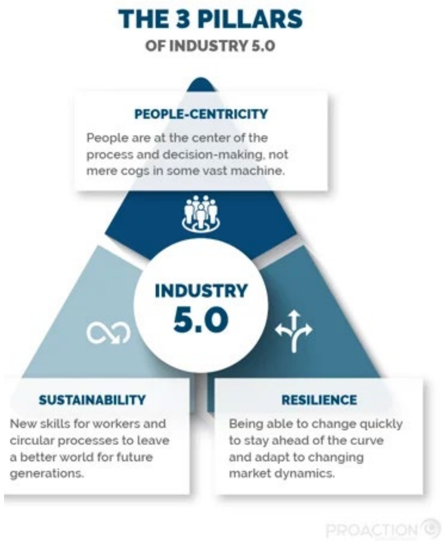
  for Sustainable and Responsible Artificial Intelligence (SRAI) technologies
- Discuss:
  - The opportunities and risks for Problem Solving with Humans and Computers
- Conclude:

  by looking forward to the Future of Technology and Society with SRAI

Historic Context: 5 Industrial Revolutions?


The Five Industrial Revolutions
RESEARCH NESTER
INDUSTRY 1.0
INDUSTRY 2.0
INDUSTRY 3.0
INDUSTRY 4.0
INDUSTRY 5.0
Mechanization and the Introduction of Steam and Water Power
Mass Production Assembly Lines Using Electrical Power
Automated production, computers, IT-systems, and robotics
The Smart Factory, autonomous systems, IoT, machine learning
Mass customization and cyber physical cognitive systems
1784
1870
1969
TODAY
TOMORROW

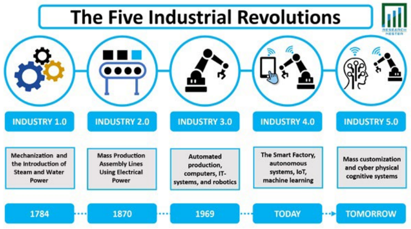
Source: Research Nester


THE 3 PILLARS
OF INDUSTRY 5.0
PEOPLE-CENTRICITY
People are at the center of the process and decision-making, not mere cogs in some vast machine.
INDUSTRY 5.0
SUSTAINABILITY
New skills for workers and circular processes to leave a better world for future generations.
RESILIENCE
Being able to change quickly to stay ahead of the curve and adapt to changing market dynamics.
PROACTION

Reference: Industry 5.0 (europa.eu)

Reference: Industry 5.0 Market Size, Share, Growth And Global Trends Analysis. 2030 (researchnester.com)

Dynamic Context: 6 Industrial Revolutions?


The Five Industrial Revolutions
RESEARCH NESTER
INDUSTRY 1.0
INDUSTRY 2.0
INDUSTRY 3.0
INDUSTRY 4.0
INDUSTRY 5.0
Mechanization and the Introduction of Steam and Water Power
Mass Production Assembly Lines Using Electrical Power
Automated production, computers, IT-systems, and robotics
The Smart Factory, autonomous systems, IoT, machine learning
Mass customization and cyber physical cognitive systems
1784
1870
1969
TODAY
TOMORROW

Source: Research Nester

The 6 Industrial Revolutions - Keywords

Reference: Peter P. Groumpos (2021) A Critical Historical and Scientific Overview of all Industrial Revolutions, IFAC-PapersOnLine, 54/13, pp 464-471, https://doi.org/10.1016/j.ifacol.2021.10.492

Reference: John Nosta (2023) The 5th Industrial Revolution: The Dawn of the Cognitive Age How technology is driving a revolution of thought. https://www.psychologytoday.com/gb/blog/the-digital-self/202310/the-5th-industrial-revolution-the-dawn-of-the-cognitive-age

Industry 1.0 (1740) *Mechanisation* (Mechanical Revolution)

Industry 2.0 (1840) *Electrification* (Electrical Revolution)

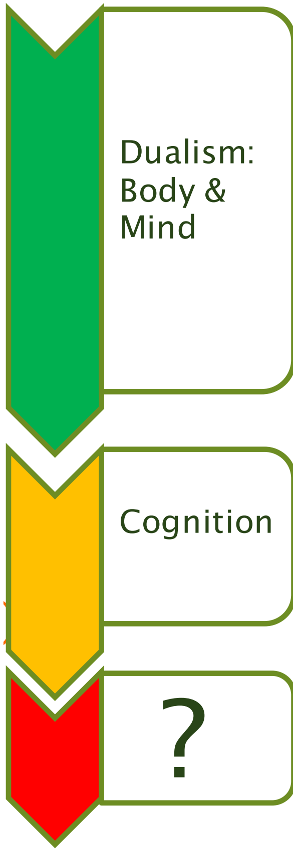
Industry 3.0 (1950) *Automation* (Automated Revolution)

Industry 3.5 (1980) *Globalisation* (Globalised Revolution)

Industry 4.0 (2000) *Digitisation/Digitalisation* (Digitise = **Data-Centred** Revolution driven by Cyber-Physical Systems, IoT, Blockchain…)

Industry 5.0 (2010) *Personalisation* (Personalised = **Human-Centred** Revolution driven by cyber-physical cognitive systems with multimodal UX)

Industry 6.0 (2020) *Humanisation* (Humanised Revolution) **Humane AI | Human-Centered Artificial Intelligence (humane-ai.eu)**



Dualism: Body & Mind

Cognition

?

What is Intelligence?

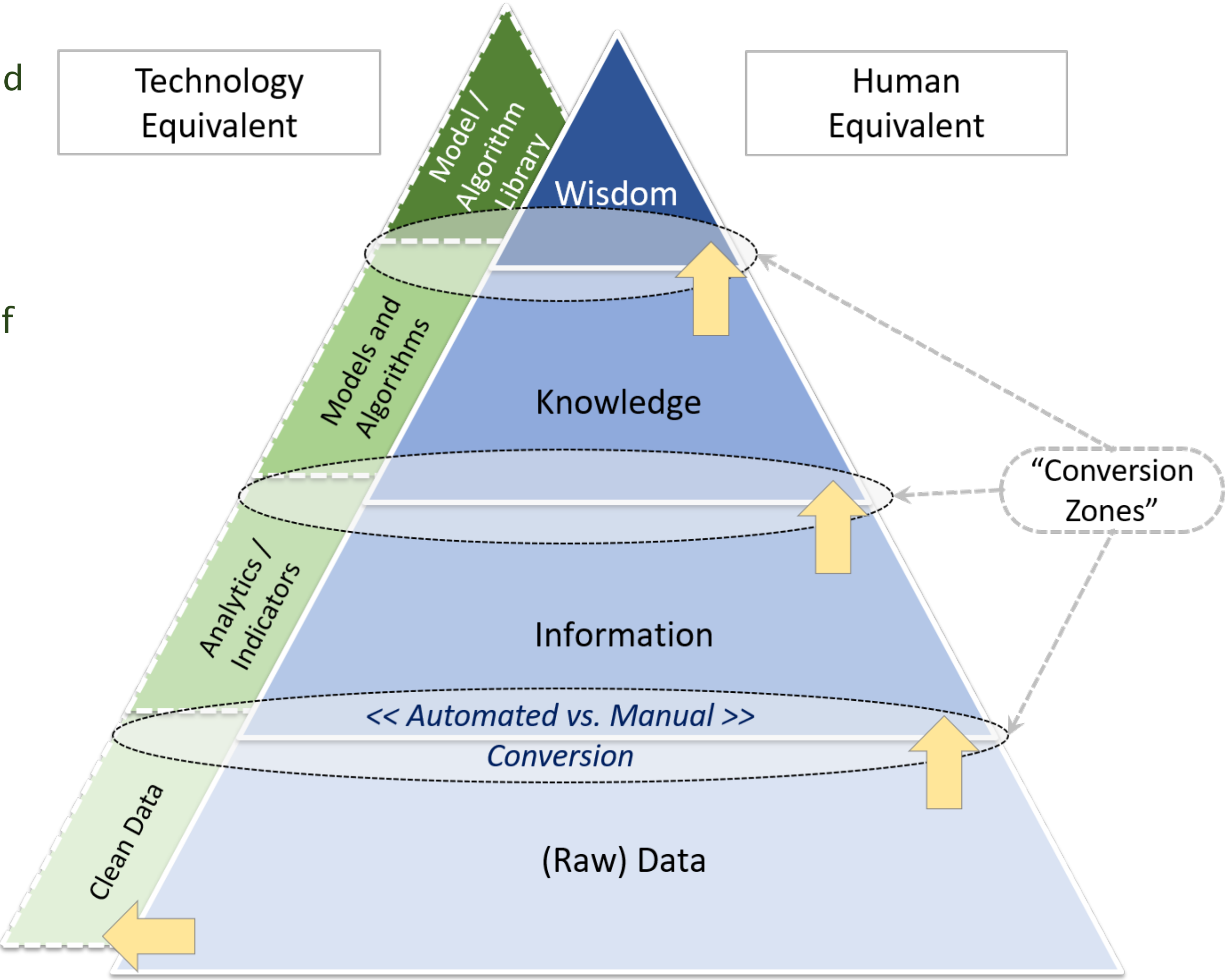
Intelligence is a complex and multifaceted ability to: **learn** (acquire; understand, process, retain) and **apply Knowledge (KDD/CRISP-DM).**

Intelligence encompasses a wide range of mental abilities and skills, allowing individuals/entities to:

- learn from experience,
- solve problems,
- reason,
- adapt to their environment,
- engage in abstract thinking and I/O (communications)

Any requirements to demonstrate Intelligence?

- I/O systems (e.g. sensors/senses)
- Memory
- Functional Brain



Source: *Alphacution*

Picture source: Paul Rowady: **DIKW hierarchy** w Automation Equivalent – Alphacution Research Conservatory

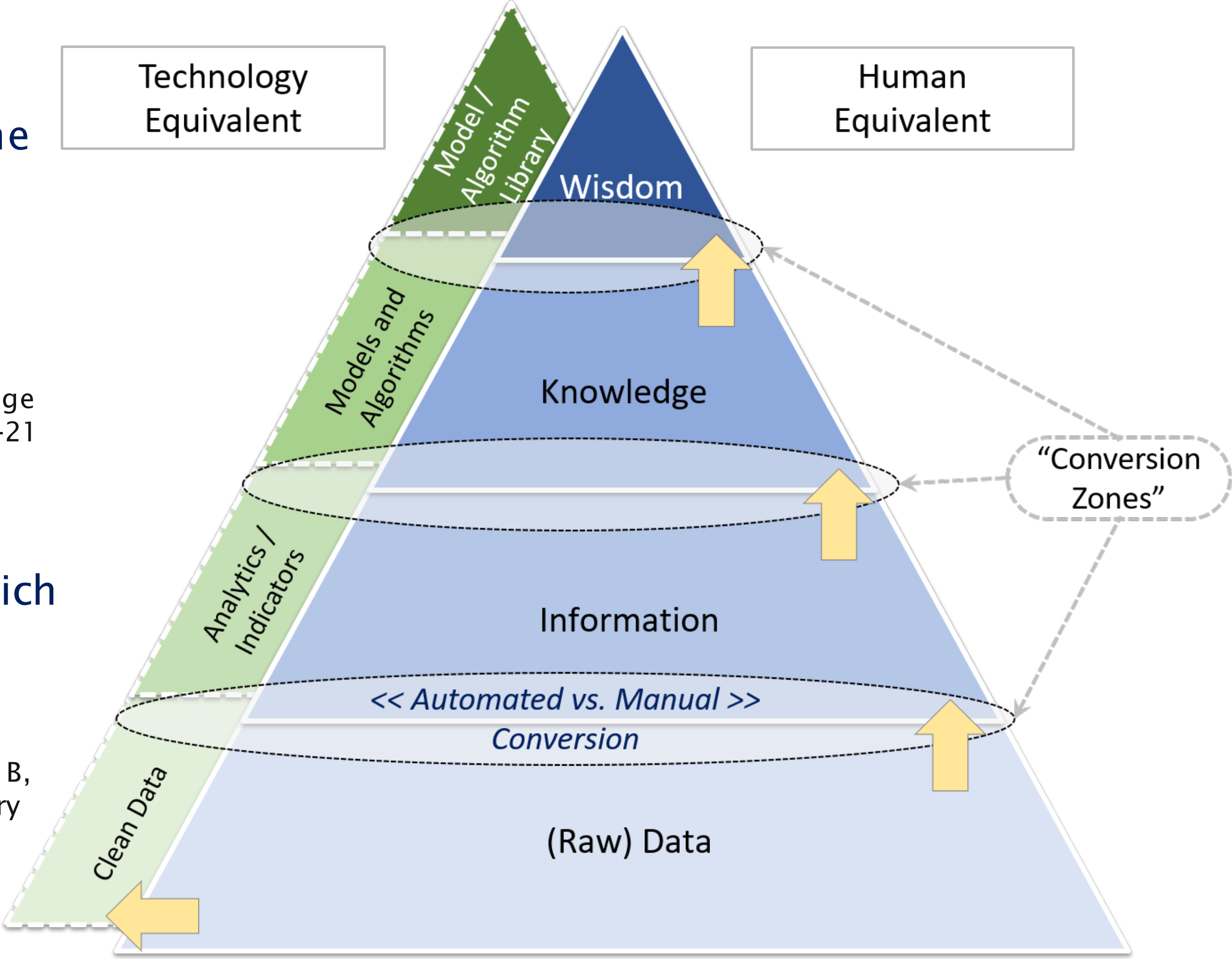
Let’s Resume: What is Intelligence?

Intelligence can be defined* as the ability to think **logically**, to conceptualise and abstract from reality.

*__Reference__: Clayton, V (1982) Wisdom and Intelligence: The Nature and Function of Knowledge in the Later Years. Int J Aging Hum Dev 15(4):315-21 doi.org/10.2190/17TQ-BW3Y-P8J4-TG40

Wisdom can be defined* as the ability to grasp human nature, which is paradoxical, contradictory, and subject to continual change.

Reference: Palanca-Castan, N, Sánchez Tajadura, B, Cofré (2021) R (2021) Towards an interdisciplinary framework about intelligence, Heliyon, 7/2, e06268, https://doi.org/10.1016/j.heliyon.2021.e06268



Source: Alphacution

Picture source: Paul Rowady: **DIKW hierarchy** w Automation Equivalent – Alphacution Research Conservatory

Reference: Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. Journal of Information Science, 35(2), 131–142. doi.org/10.1177/0165551508094050

DIKW (cont’d): more on Knowledge

Data (facts and figures) are (raw input) Information.

However, there is Information that is not Data.

Know-that is recordable information (including processed data).

Knowledge is actually know-how (building on know-what).

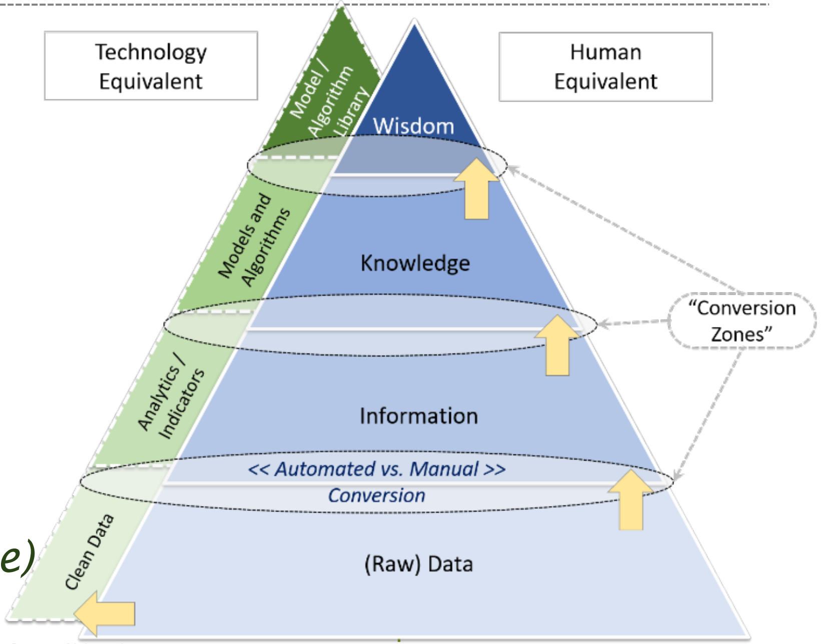
Information science uses a propositional account of knowledge i.e. **knowledge-that (weak knowledge)**.

This makes *know-that (that is articulable and recordable) and information synonymous*.

Knowledge is the theoretical and practical comprehension of a certain domain, that supports making decisions.

DIKW and epistemology see **knowledge as know-how**; and, in turn, this tends to **make knowledge inarticulable and not recordable.**

Reference: Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. Journal of Information Science, 35(2), 131–142. doi.org/10.1177/0165551508094050



Source: Alphacution

Picture source: Paul Rowady: **DIKW hierarchy** w Automation Equivalent – Alphacution Research Conservatory

?	Philosophy (Science of Wisdom)
Knowledge Engineering	Epistemology
Information Engineering	Information Science
Data Engineering	Data Science

DIKW (cont’d): more on Knowledge

Data is (raw) Information.

However, there is Information that is not Data.

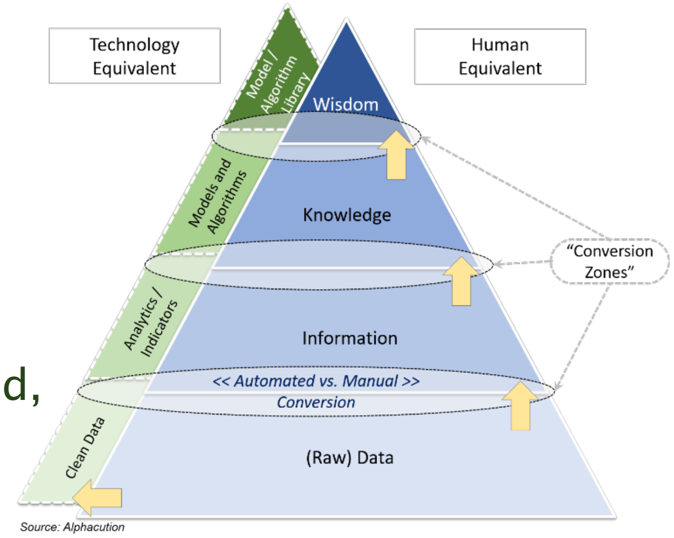
Know-that is recordable information.

Knowledge is actually know-how.

Knowledge is the theoretical and practical comprehension of a certain domain, that supports making decisions.

DIKW and epistemology see **knowledge as know-how**; and, in turn, this tends to **make knowledge inarticulable and not recordable.**

Reference: Frické, M. (2009). The knowledge pyramid: a critique of the DIKW hierarchy. Journal of Information Science, 35(2), 131–142. doi.org/10.1177/0165551508094050



Source: Alphacution

Picture source: Paul Rowady: **DIKW hierarchy** w Automation Equivalent – Alphacution Research Conservatory

Epistemology = the philosophical study of the nature, origin, and limits of human knowledge. The term is derived from the Greek epistēmē (“knowledge”) and logos (“reason”), and accordingly the field is sometimes referred to as the theory of knowledge. Along with metaphysics, logic, and ethics, it is one of the four main branches of philosophy.

Reference: https://www.britannica.com/topic/epistemology

?	Philosophy (Science of Wisdom)
Knowledge Engineering	Epistemology
Information Engineering	Information Science
Data Engineering	Data Science

Key Definitions

Reference: Updates to the OECD's definition of an AI system explained – OECD.AI
Reference: EU AI Act [The Act Texts | EU Artificial Intelligence Act]
Reference: Data science and AI glossary | The Alan Turing Institute

AI System ('artificial intelligence system') means software that is developed with one or more of the techniques and approaches listed in Annex I (EU AI Act) and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.

Foundation Model = A machine learning model trained on a vast amount of (big) data so that it can be easily adapted for a wide range of applications. A common type of foundation model is large language models, which power chatbots such as ChatGPT.

Large Language Model (LLM) = A type of foundation model that is trained on a vast amount of textual data in order to carry out language-related tasks. Large language models power the new generation of chatbots, and can generate text that is indistinguishable from human-written text. They are part of a broader field of research called natural language processing, and are typically much simpler in design than smaller, more traditional language models.

Key Definitions

Reference: Updates to the OECD's definition of an AI system explained – OECD.AI
Reference: EU AI Act [The Act Texts | EU Artificial Intelligence Act]
Reference: Data science and AI glossary | The Alan Turing Institute

Generative Adversarial Network = A machine learning technique that can generate data, such as realistic 'deepfake' images, which is difficult to distinguish from the data it is trained on. A GAN is made up of two competing elements: a generator and a discriminator. The generator creates fake data, which the discriminator compares to real 'training' data and feeds back with where it has detected differences. Over time, the generator learns to create more realistic data, until the discriminator can no longer tell what is real and what is fake.

Generative AI = An artificial intelligence system that generates text, images, audio, video or other media in response to user prompts. It uses machine learning techniques to create new data that has similar characteristics to the data it was trained on (see 'generative adversarial network'), resulting in outputs that are often indistinguishable from human-created media (see 'deepfake').

Chatbot = A software application that has been designed to mimic human conversation, allowing it to talk to users via text or speech. Previously used mostly as virtual assistants in customer service, chatbots are becoming increasingly powerful and can now answer users' questions across a variety of topics, as well as generating stories, articles, poems and more (see also 'Generative AI').

Key Definitions: Artificial General Intelligence

General AI/ Artificial General Intelligence (AGI, The Singularity): represents a theoretical form of artificial intelligence (AI) that could solve any task using human-like cognitive abilities. AGI aims to perform as well as or better than humans across a wide range of cognitive functions. The exact definition of AGI is still debated: modern large language models (LLMs) like GPT-4o, CoPilot and Gemini are early, incomplete (industry 4.0) forms of AGI still able to pass some (Turing) Tests. In science fiction and futures studies, AGI is a common topic, and there is contention over its potential impact on humanity (AI risks).

Reference: https://plato.stanford.edu/entries/artificial-intelligence/#StroVersWeakAI

“Weak” AI or “Narrow” AI seeks to build machines that appear to outperform human persons for a dedicated purpose or specific task. It turns relevant big data in usable information: Apple's Siri, Amazon's Alexa, IBM watsonx™, self-driving vehicles.

“Strong” AI is a theoretical form of AGI where the machine would hypothetically possess human-level intelligence; it would be self-aware including phenomenal consciousness, and it would have the ability to solve problems, learn, and plan for the future.

“Strong” AI can also be defined as the form of AI that aims at a system able to pass not just the Turing Test (again, abbreviated as TT), but the Total Turing Test (Harnad 1991), showing more than linguistic indistinguishability, for example the superhuman and rogue computer assistant in 2001: A Space Odyssey.

Key Definitions: The Singularity

General AI/ Artificial General Intelligence (AGI, The Singularity): represents a theoretical form of artificial intelligence (AI) that could solve any task using human-like cognitive abilities. AGI aims to perform as well as or better than humans across a wide range of cognitive functions. The exact definition of AGI is still debated: **modern large language models (LLMs)** like GPT-4, CoPilot and Gemini **are early, incomplete forms of AGI**. In science fiction and futures studies, AGI is a common topic, and there is contention over its potential impact on humanity (AI risks).

Reference: https://plato.stanford.edu/entries/artificial-intelligence/#StroVersWeakAI (?)

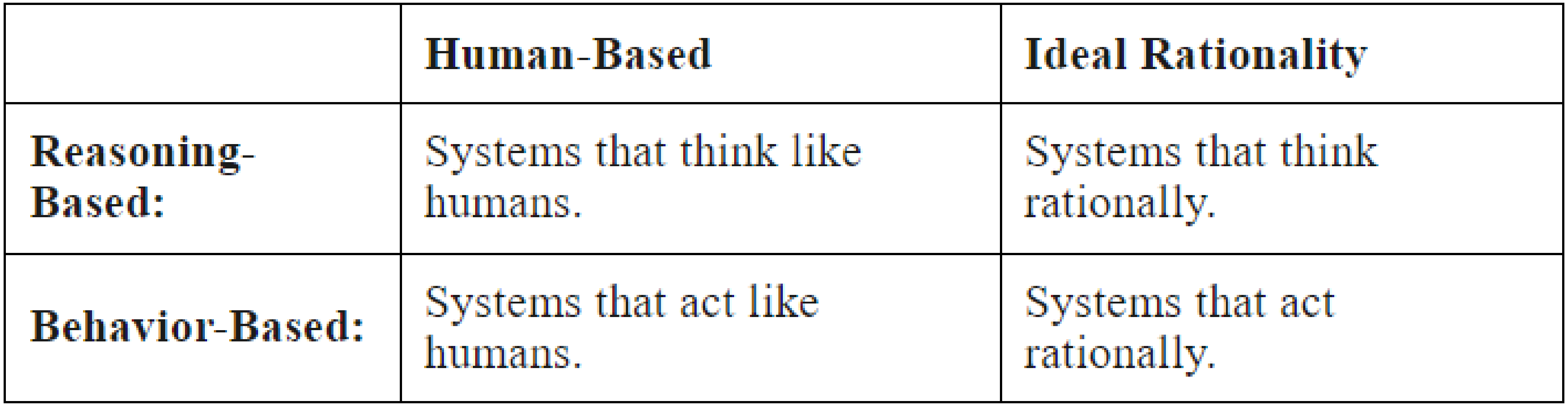
"Weak" AI or "Narrow" AI seeks to build machines that appear to outperform human persons for a dedicated purpose or specific task. It turns relevant big data in usable information: Apple's Siri, Amazon's Alexa, IBM watsonx™, self-driving vehicles.

"Strong" AI is a theoretical form of AGI where the machine would hypothetically possess human-level intelligence; it would be self-aware including phenomenal consciousness, and it would have the ability to solve problems, learn, and plan for the future.

"Strong" AI can also be defined as the form of AI that aims at a system able to pass not just the Turing Test (again, abbreviated as TT), but the Total Turing Test (Harnad 1991), showing more than linguistic indistinguishability, for example the superhuman and rogue computer assistant in 2001: A Space Odyssey.

Key Definitions: Intelligent Systems

Reference: Russel & Norvig – AI: a Modern Approach (AIMA):



	Human-Based	**Ideal Rationality**
Reasoning-Based:	Systems that think like humans.	Systems that think rationally.
Behavior-Based:	Systems that act like humans.	Systems that act rationally.

Four Possible Goals for AI According to AIMA

The 5 Ethical Principles

Responsible AI or Ethical AI or Trustworthy AI? or…
– interchangeability of these attributes opens potential concerns and challenges!!!

The global landscape of **AI ethics** guidelines shows that there is a global convergence around **five ethical principles**: **Transparency, Impartiality** (Justice, **Fairness**, Non-Maleficence), Reliability/Robustness, Accountability/**Responsibility**, and **Privacy.**

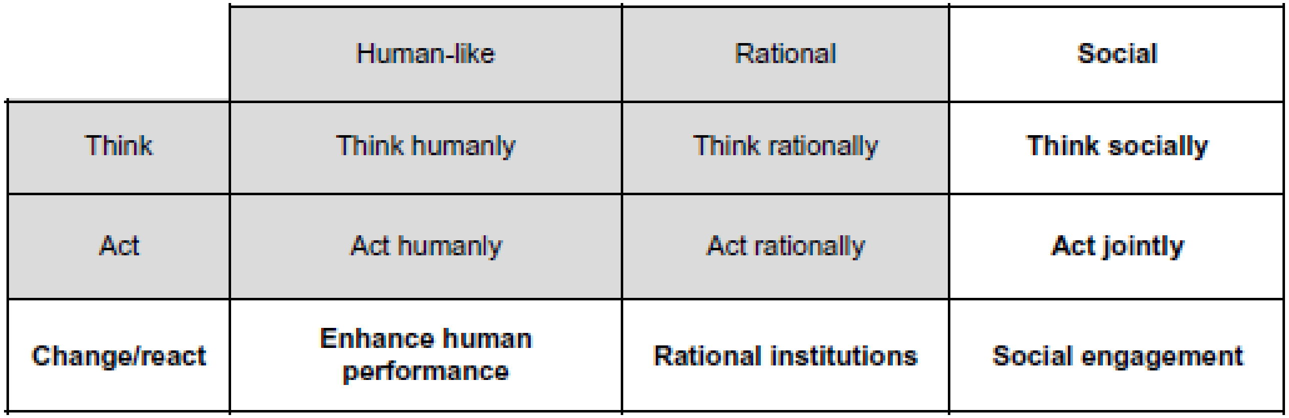
It is not the AI artefact or application that needs to be ethical, trustworthy, or responsible. Rather, it is the social component of this ecosystem that can and should take responsibility and act in consideration of an ethical framework such that the overall system can be trusted by the society:

References:

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.8cd550d1

Dignum, V. Responsible Artificial Intelligence – from Principles to Practice 2205.10785v1.pdf (arxiv.org) ACM SIGIR Forum

Anna Jobin, Marcello Ienca, and Effy Vayena (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9):389-399.



	Human-like	Rational	**Social**
Think	Think humanly	Think rationally	**Think socially**
Act	Act humanly	Act rationally	**Act jointly**
Change/react	**Enhance human performance**	**Rational institutions**	**Social engagement**

Fig. 1 Social perspectives in AI

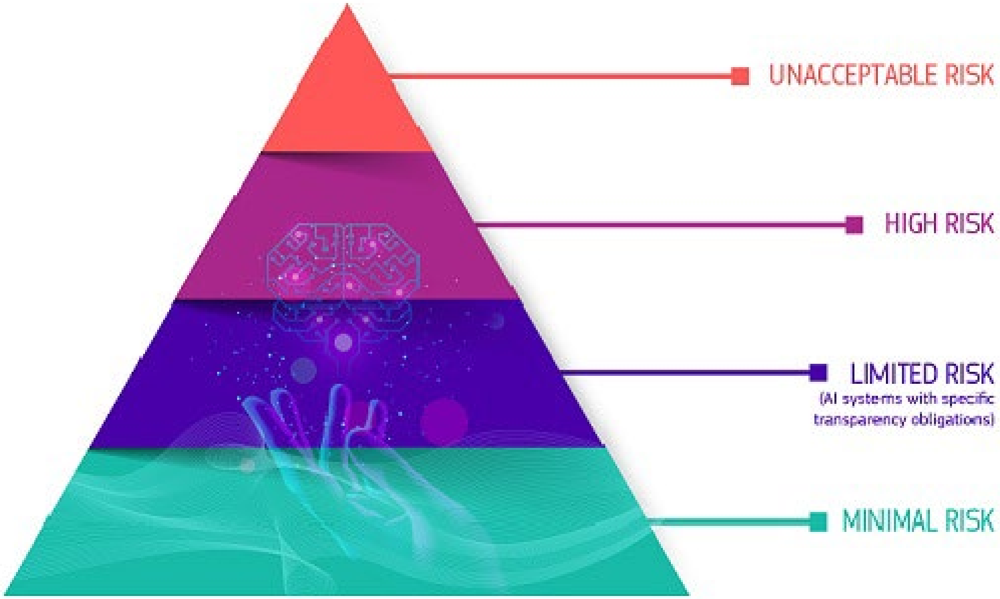
Key Definitions: High Risk in AI Systems

References:

EU AI Act [The Act Texts | EU Artificial Intelligence Act]

EU AI Act – a risk based approach

EU AI Office



High-Risk AI System: if both conditions are fulfilled:

(a) the AI system is intended to be used as a safety component of a product, or is itself a product, *covered by the Union harmonisation legislation listed in Annex II*;

(b) the product whose safety component is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product *pursuant to the Union harmonisation legislation listed in Annex II.*

High Risk AI systems used in: critical infrastructures (life/health risks e.g. transport, surgeries); educational, employment; essential private and public services (e.g. credit); law enforcement; visa.

Limited risk refers to the risks associated with lack of **transparency** in AI usage: AI-generated content (**chatbots**, **deep fakes**).

Minimal or no risk: AI-enabled *video games* or spam filters

What is Responsibility?

Responsibility refers to the duty or obligation of an individual or group to fulfil certain roles, tasks, or duties in a reliable and accountable manner.

Responsibility is an important aspect of **ethical and moral behaviour**, as it requires individuals to take ownership of their actions and acknowledge the impact they have on themselves, others, and the wider community.

Key aspects of Responsibility include:

Accountability: recognising and accepting the consequences of one's actions.

Reliability: trustworthy in fulfilling commitments, obligations and meeting expectations.

Ethical Decision-making: avoiding actions that harm others or violate societal norms.

Learning from Mistakes: using them as opportunities for growth and improvement.

Responsibility plays a significant role in **building trust and respect** in relationships and contributes to a sense of **integrity and dignity** in individuals and communities.

What is Ethical AI?

EAI is the duty, obligation or expectation from technology (designers, owners or users) to provide or demonstrate features of:

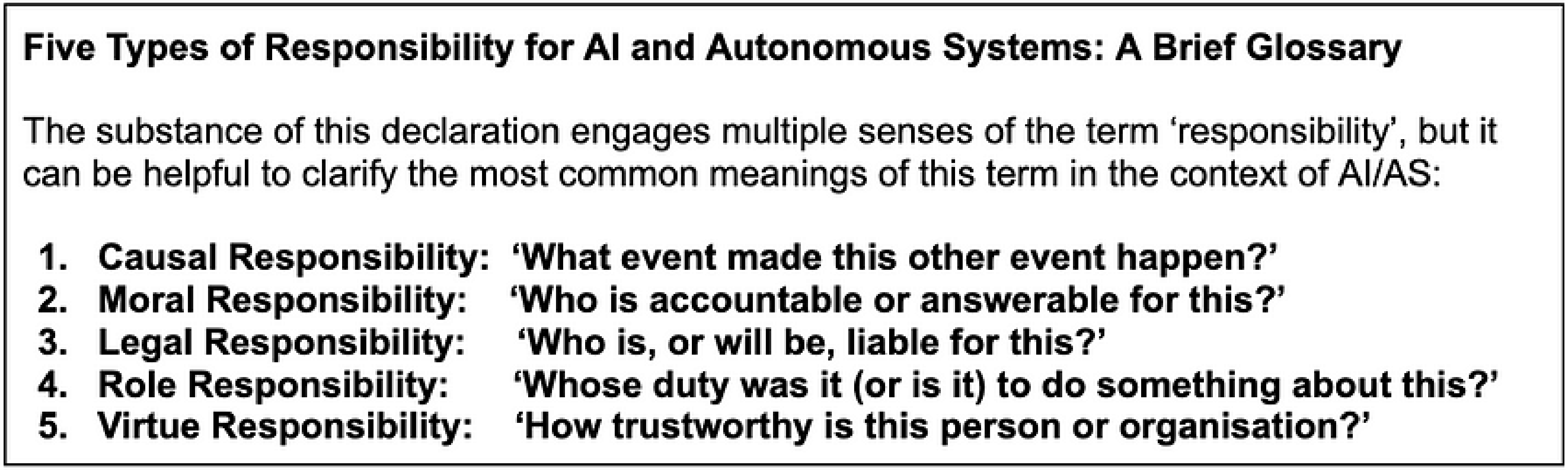
ACCOUNTABILITY
AWARENESS
BALANCE
CONSCIENCE
CONSIDERATION
EXPLAINABILITY
EFFECTIVENESS
EFFICIENCY
EMPATHY
ETHICS
FAIRNESS
RELIABILITY
RESPONSIBILITY
ROBUSTNESS
SAFETY
SUSTAINABILITY
TRANSPARENCY
TRUSTWORTHINESS
VISION
UNBIAS

The 5 RAI Pillars

What is Responsible AI? | IBM: includes the following **RAI Pillars**: 1) **Explainability** (with the **Principles**: Prediction accuracy; Traceability; Decision understanding); 2) **Fairness** (Principles: Diverse and representative data; Bias-aware algorithms; Diverse development teams; Ethical AI review boards); 3) **Robustness**; 4) **Transparency**; 5) **Privacy**

https://www.kolena.com/blog/7-pillars-of-responsible-ai#7-pillars-of-responsible-ai: Accountability; **Transparency**; **Explainability; Interpretability**; **Fairness**; Unbias; **Privacy** Protection; Security; Resilience; **Validity; Reliability**; Safety

Edinburgh Declaration on Responsibility for Responsible AI:



Five Types of Responsibility for AI and Autonomous Systems: A Brief Glossary

The substance of this declaration engages multiple senses of the term ‘responsibility’, but it can be helpful to clarify the most common meanings of this term in the context of AI/AS:

1. **Causal Responsibility:** **‘What event made this other event happen?’**
2. **Moral Responsibility:** **‘Who is accountable or answerable for this?’**
3. **Legal Responsibility:** **‘Who is, or will be, liable for this?’**
4. **Role Responsibility:** **‘Whose duty was it (or is it) to do something about this?’**
5. **Virtue Responsibility:** **‘How trustworthy is this person or organisation?’**

What is eXplainable Intelligence?

Reference: Four Principles of Explainable Artificial Intelligence (National Institute of Standards and Technology, US Department of Commerce)

Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.

Meaningfulness: A system provides explanations that are understandable to the intended consumer(s).

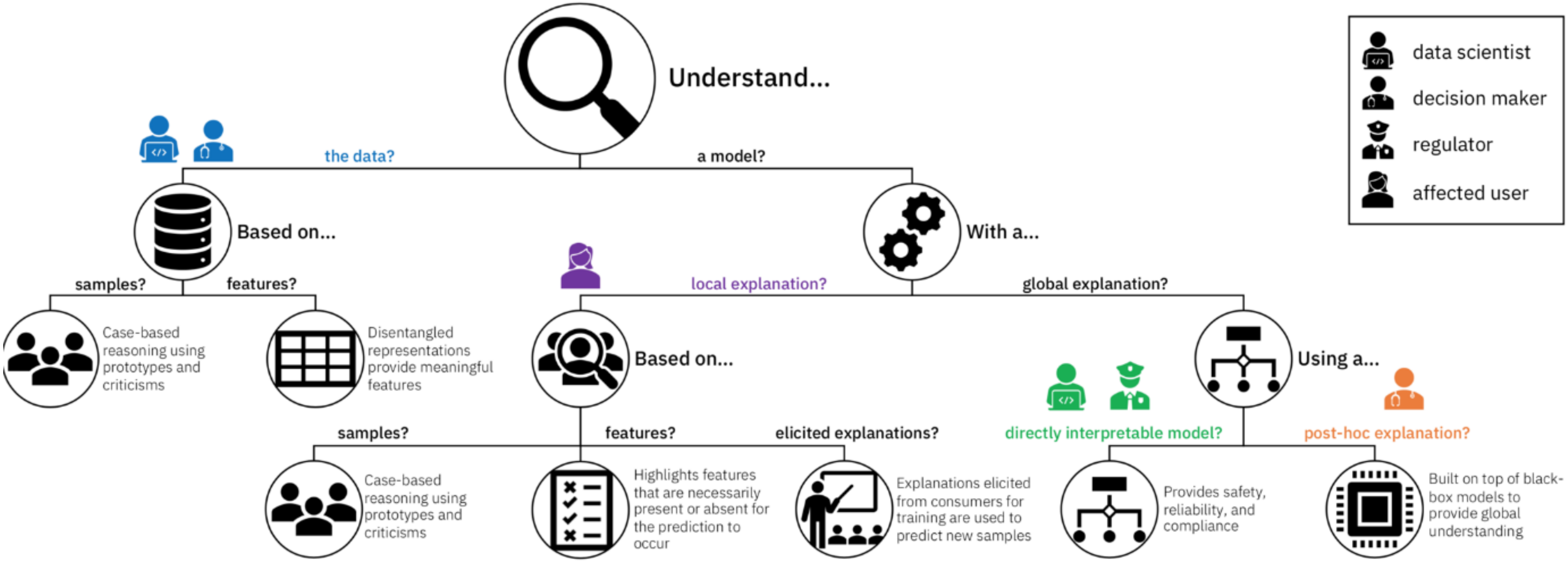
Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.

Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

What is eXplainable Intelligence?

IBM AI Explainability 360 toolkit: Explainability is not a singular approach. There are many ways to explain how machine learning makes predictions, including: data vs. model; **directly interpretable vs. post hoc explanation**; local vs. global; static vs. interactive:

We can see IBM uses interchangeably on subdomains of reasoning Interpretability, Explainability and Understanding.


Understand...
the data?
a model?
Based on...
samples?
features?
Case-based reasoning using prototypes and criticisms
Disentangled representations provide meaningful features
With a...
local explanation?
global explanation?
Based on...
samples?
features?
elicited explanations?
Case-based reasoning using prototypes and criticisms
Highlights features that are necessarily present or absent for the prediction to occur
Explanations elicited from consumers for training are used to predict new samples
Using a...
directly interpretable model?
post-hoc explanation?
Provides safety, reliability, and compliance
Built on top of black-box models to provide global understanding
data scientist
decision maker
regulator
affected user

eXplainability vs Interpretability of AI

Reference: Psychological Foundations of Explainability and Interpretability in AI with examples (National Institute of Standards and Technology, US Department of Commerce)

Interpretability is the degree to which an observer can understand the cause of a decision. It is the success rate that humans can predict for the result of an AI output, while **eXplainability** goes a step further and looks at how the AI arrived at the result.

Interpretable models should provide users with a description of what a stimulus, such as a datapoint or model output, means in context.

- An **explanation** seeks to describe the process that generated an output.
- Thus, an **explanation of an algorithm's output is justified relative to an implementation, or technical process**, that was used to generate a specific output.
- In contrast, an **interpretation is justified relative to the functional purpose of the algorithm.**

eXplainability vs Interpretability of AI

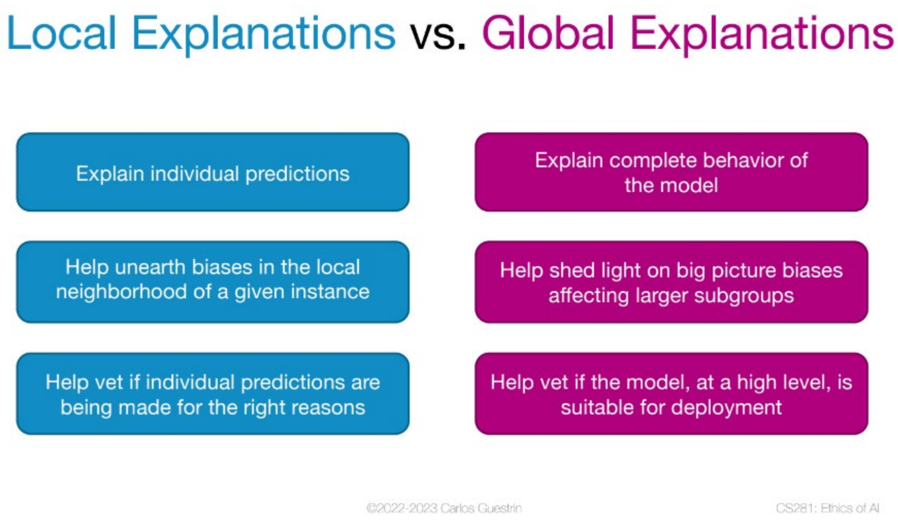
Reference: Interpretable Models vs Post-hoc Explanations:
CS281: Ethics of Artificial Intelligence - Stanford University

Ante-hoc eXplainability (sometimes also called ***intrinsic Interpretability*** or ***Transparent model design***) is the strategy of directly training explainable models.

Post-hoc eXplainability: is the strategy of explaining a (plausibly opaque) model after it was trained.

- model-agnostic eXplainability methods: are the methods that work independently of the underlying model;

- model-specific eXplainability methods: are the methods that only work for certain models or model classes.



Local Explanations vs. Global Explanations

Local Explanations	Global Explanations
Explain individual predictions	Explain complete behavior of the model
Help unearth biases in the local neighborhood of a given instance	Help shed light on big picture biases affecting larger subgroups
Help vet if individual predictions are being made for the right reasons	Help vet if the model, at a high level, is suitable for deployment

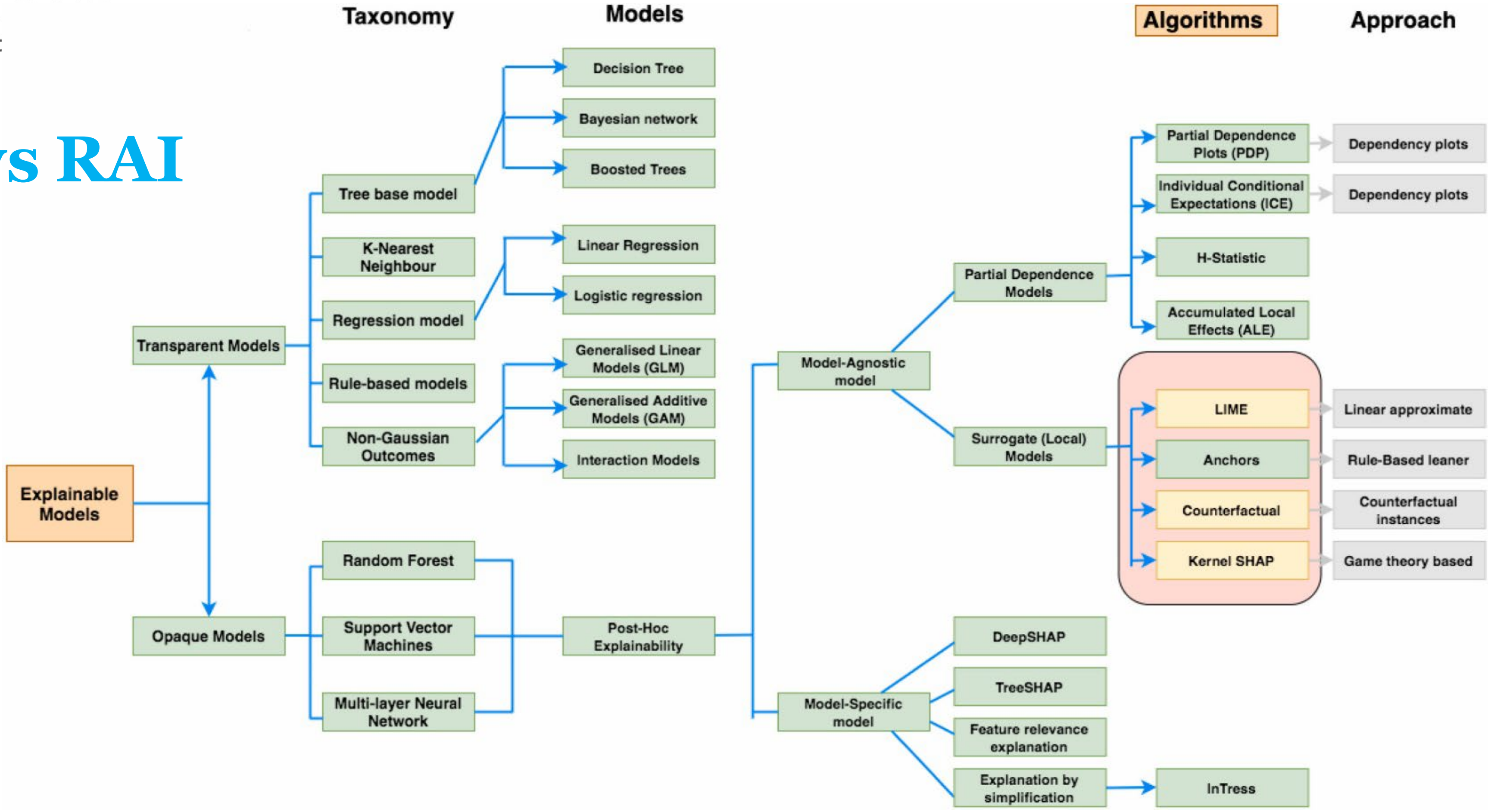
©2022-2023 Carlos Guestrin

CS281: Ethics of AI

UNIVERSITY of BRADFORD

School of Comput
AI & Electronics

XAI vs RAI



Reference: Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications

How does eXplainable AI relate to Responsible AI?

Explainable AI and Responsible AI have similar objectives, yet different approaches. Main differences between XAI - RAI:

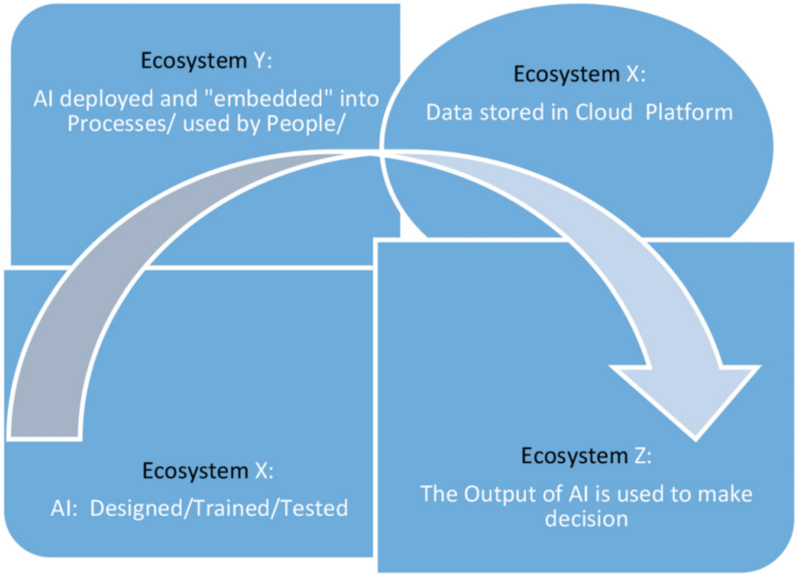
1. Explainable AI looks at AI results after the results are computed.
2. Responsible AI - during the planning stages makes the AI algorithm responsible before the results are computed.
3. Explainable and Responsible AI can work together to make better AI.

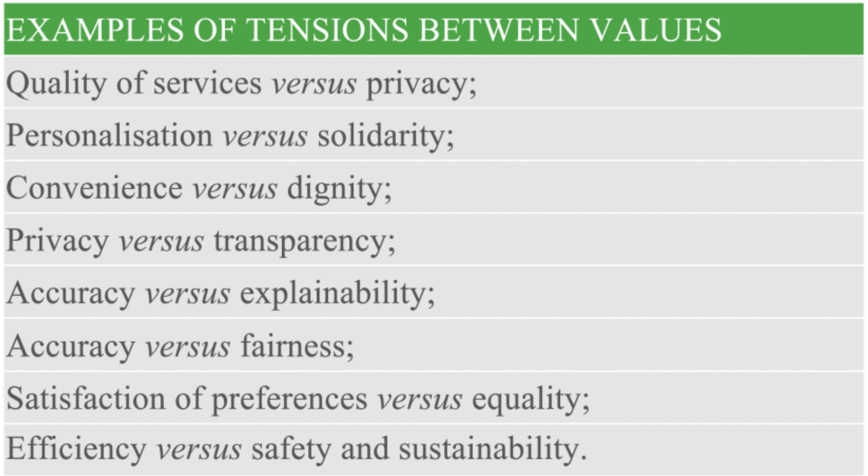
Trustworthy AI

Reference: Z-Inspection®: A Process to Assess Trustworthy AI | IEEE Journals & Magazine | IEEE Xplore

The 7 requirements established by the EU High-Level Experts Group (HLEG) on AI's Guidelines for **Trustworthy AI**:

1. human agency and oversight;
2. technical robustness and safety;
3. privacy and data governance;
4. transparency;
5. diversity, non-discrimination, and fairness;
6. societal and environmental well-being;
7. accountability.


Ecosystem Y:
AI deployed and "embedded" into Processes/ used by People/
Ecosystem X:
Data stored in Cloud Platform
Ecosystem X:
AI: Designed/Trained/Tested
Ecosystem Z:
The Output of AI is used to make decision



EXAMPLES OF TENSIONS BETWEEN VALUES
Quality of services *versus* privacy;
Personalisation *versus* solidarity;
Convenience *versus* dignity;
Privacy *versus* transparency;
Accuracy *versus* explainability;
Accuracy *versus* fairness;
Satisfaction of preferences *versus* equality;
Efficiency *versus* safety and sustainability.

Transparent AI

Reference: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI | Journal of Medical Ethics (bmj.com)

Transparency is an **epistemic** manoeuvre intended to offer reasons to believe that certain algorithmic procedures render a reliable output. Furthermore, according to the partisan of transparency, such a belief also entails that the output of the algorithm is interpretable by humans.

- According to **epistemic opacity**, humans are neither able to account for the state of the algorithm (i.e. its variables, relations, system status, etc) previous to the halt, nor to predict any of the future state of the algorithm after the halt. Furthermore, humans would not be able to account for the state of the algorithm and its variables at the time of the halt either. The implications of a **fully epistemically opaque algorithm** are that medical AI work as truly obscure entities of which very little can be epistemically warranted.
- **Methodological opacity** stems from the complexities inherent to the design and programming of algorithms.

Black box algorithms are methodologically and epistemically opaque systems.

Interpretability in Machine Learning (ML) is therefore giving humans a mental model of the machine (ML) model behaviour.

Transparent AI Case Study: AI Models in Medicine

Generally, input data sets are highly dimensional and complex, and therefore needing specialised learning algorithms,, ML models are also complex, therefore opaque and not transparent, incomprehensible for the human user, although training and testing results could be decisively good.

From an epistemologic viewpoint, the transparency of AI models in medicine refer tc the understanding of the nature, origin of the output result, objective of modelling justification, transparency and trust, as well as their capacity to explain their output.

Applications: medical scans/image processing cancer identification, and its type(s), prioritisation of patients and decisions.

In such cases, even if used just as complementary tools to support medical experts in their decisions, black box models dc not offer information to support the fina decision explicitly.

Conclusion: GenAI won’t take radiologists’ jobs, but radiologists supported by Trustworthy AI models will take jobs of radiologists without knowledge of GenAI (Andrew Ng, Generative AI for Everyone).

Reference: Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI | Journal of Medical Ethics (bmj.com)

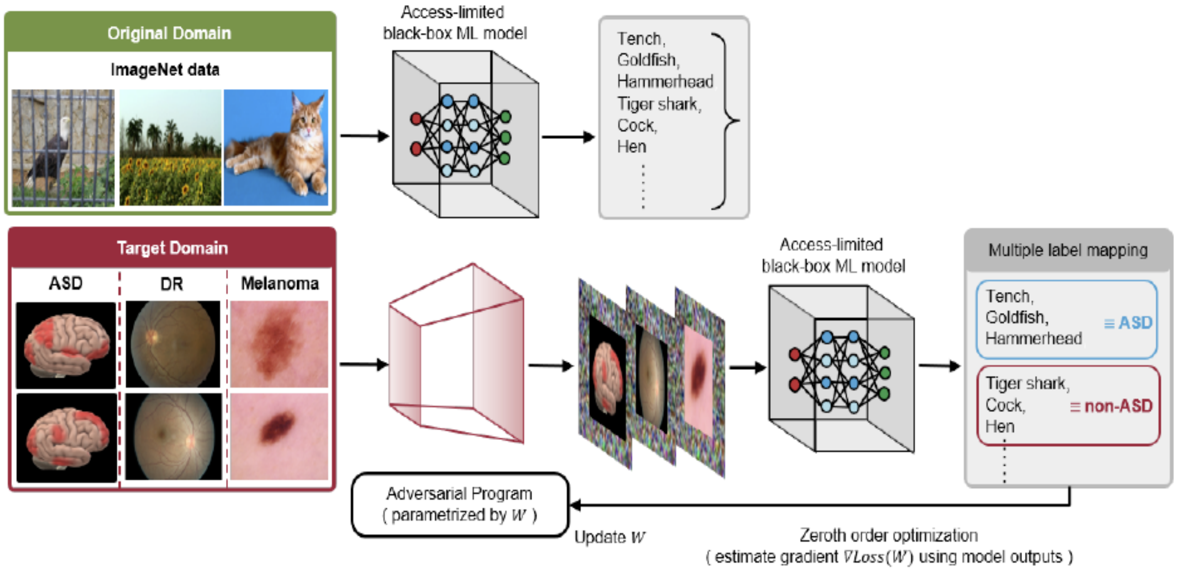


Figure 1. Schematic overview of our proposed black-box adversarial reprogramming (BAR) method.

Transfer Learning without Knowing: Reprogramming Black-box Machine Learning Models with Scarce Data and Limited Resources, Yun-Yun Tsai, Pin-Yu Chen, Tsung-Yi Ho, in Proceeding of International Conference on Machine Learning (ICML), 2020

What is Fairness? Bias? Discrimination?

References:

What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective | ACM Computing Surveys

A Survey on Bias and Fairness in Machine Learning | ACM Computing Surveys

Fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.

An **unfair algorithm** is one whose decisions are skewed toward a particular group of people because of:

A. Bias: can be considered as a source for unfairness that is due to the data collection, sampling, and measurement.

B. Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally.

C. Fairness by Design can be approached as **Fairness Through Awareness**

- it opens the opportunity of the discussion of Humanised AI and Conscience.

Measuring Bias in Algorithms?

References:

What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective | ACM Computing Surveys

A Survey on Bias and Fairness in Machine Learning | ACM Computing Surveys

Bias as a source for unfairness due to the data collection, sampling, and measurement – opens statistical opportunities for quantitative approaches.

A. Unfair Algorithms are classified **by the Types of Bias**:

A.1. Biases from Data in Algorithm

A.2. Bias in Algorithm to User

A.3. Bias in User to Data Process

B. Measuring Discrimination in Algorithms?

References:

What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective | ACM Computing Surveys

A Survey on Bias and Fairness in Machine Learning | ACM Computing Surveys

Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally.

B. Unfair Algorithms are classified **by the Types of Discrimination**:

B.1. Explainable Discrimination

B.2. Unexplainable Discrimination

B.3. Sources of Discrimination

C. Fairness by Design

References:

What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective | ACM Computing Surveys

A Survey on Bias and Fairness in Machine Learning | ACM Computing Surveys

C. Fairness by Design can be approached as **Fairness Through Awareness**

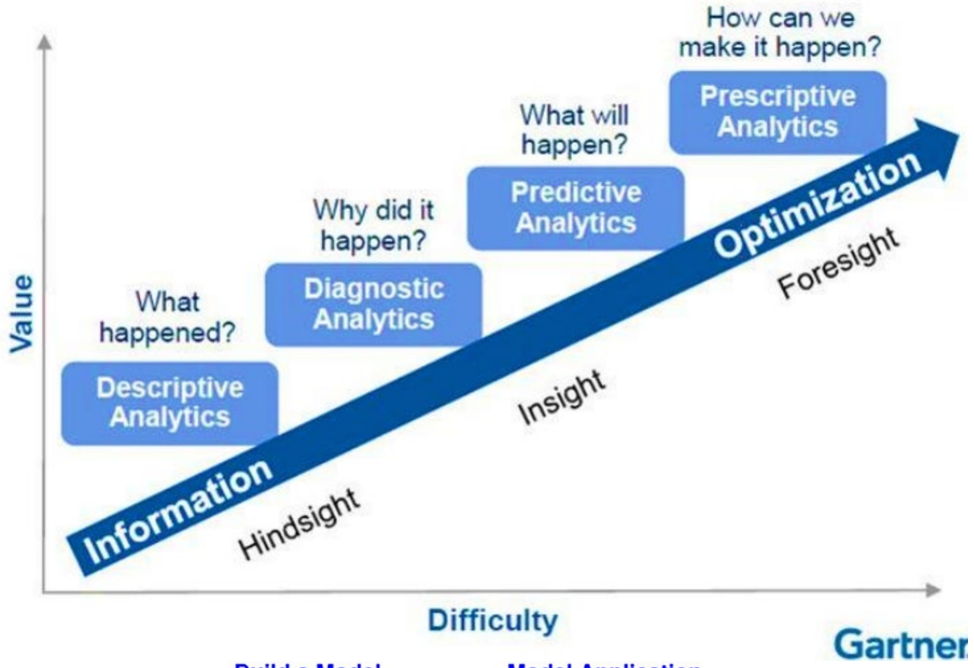
C.1. Fairness Through Awareness: "An algorithm is fair if it gives similar predictions to similar individuals". In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.

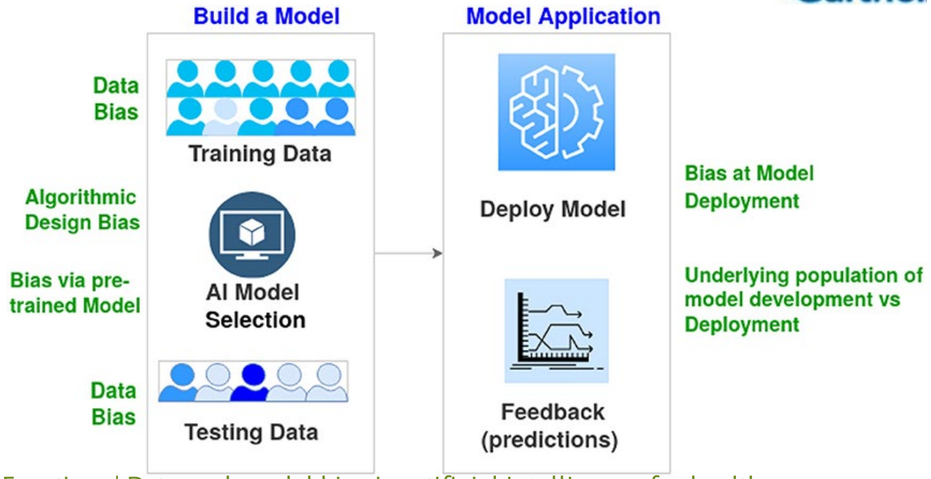
C.2. Fairness Through Unawareness: "An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process".

C.3. Treatment Equality. "Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories".

What challenges are brought by AI?

Data-centric AI systems build on **Empirical Knowledge** = the knowledge based on experience and directly observed (observation, experimentation, induction) and is the basis of ***a posteriori* knowledge**. The opposite of the empirical knowledge is ***a priori* knowledge** (based on logical deduction, reason in expert systems, rule-based systems). Are GenAI hallucinations then a stepping point for anticipation in AI systems? (**reference**: Rosen's Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations | SpringerLink theory that has influenced the turn towards anticipation in foresight and governance studies).



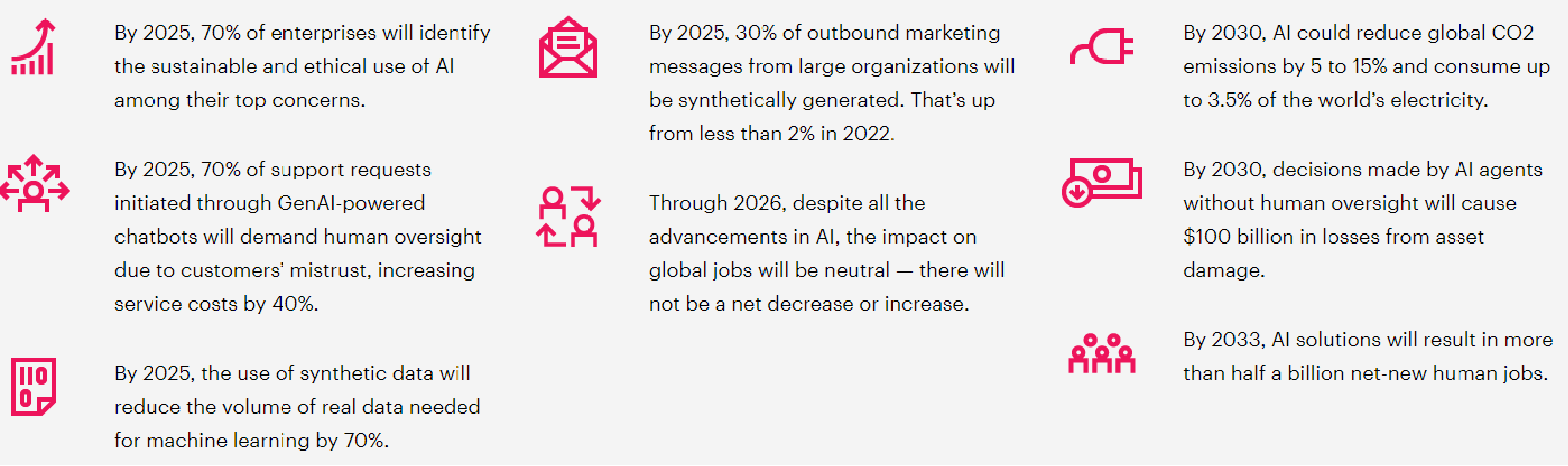


Frontiers | Data and model bias in artificial intelligence for healthcare applications in New Zealand (frontiersin.org)

Gartner Predictions about Generative AI

Reference: What Generative AI Means for Business (gartner.com)

GenAI will greatly impact product development, customer experience, employee productivity and innovation.



By 2025, 70% of enterprises will identify the sustainable and ethical use of AI among their top concerns.

By 2025, 70% of support requests initiated through GenAI-powered chatbots will demand human oversight due to customers' mistrust, increasing service costs by 40%.

By 2025, the use of synthetic data will reduce the volume of real data needed for machine learning by 70%.

By 2025, 30% of outbound marketing messages from large organizations will be synthetically generated. That's up from less than 2% in 2022.

Through 2026, despite all the advancements in AI, the impact on global jobs will be neutral — there will not be a net decrease or increase.

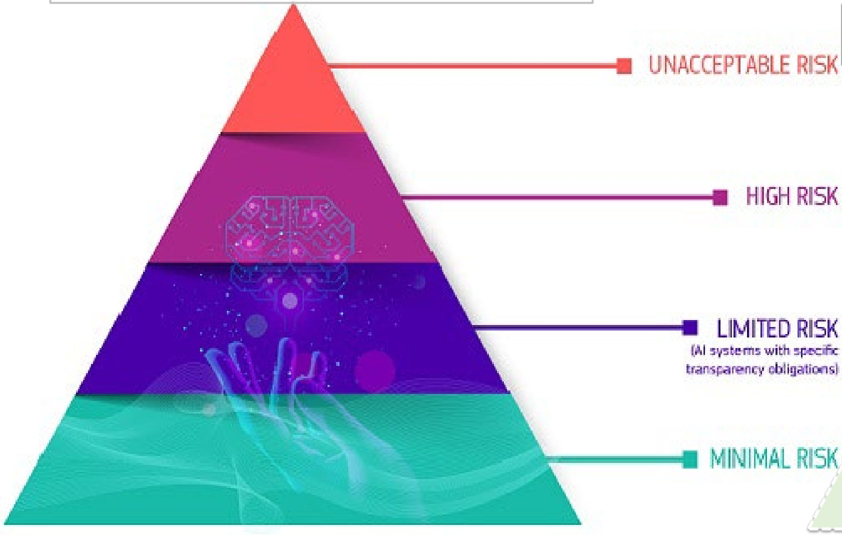
By 2030, AI could reduce global CO2 emissions by 5 to 15% and consume up to 3.5% of the world's electricity.

By 2030, decisions made by AI agents without human oversight will cause $100 billion in losses from asset damage.

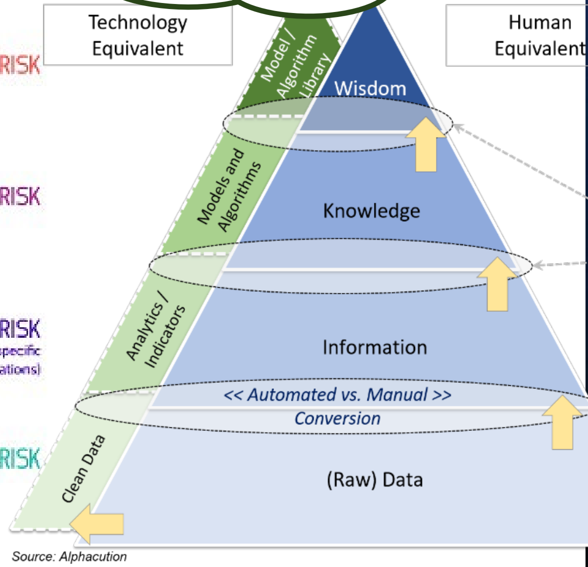
By 2033, AI solutions will result in more than half a billion net-new human jobs.

High Risk in AI Systems: the AGI?

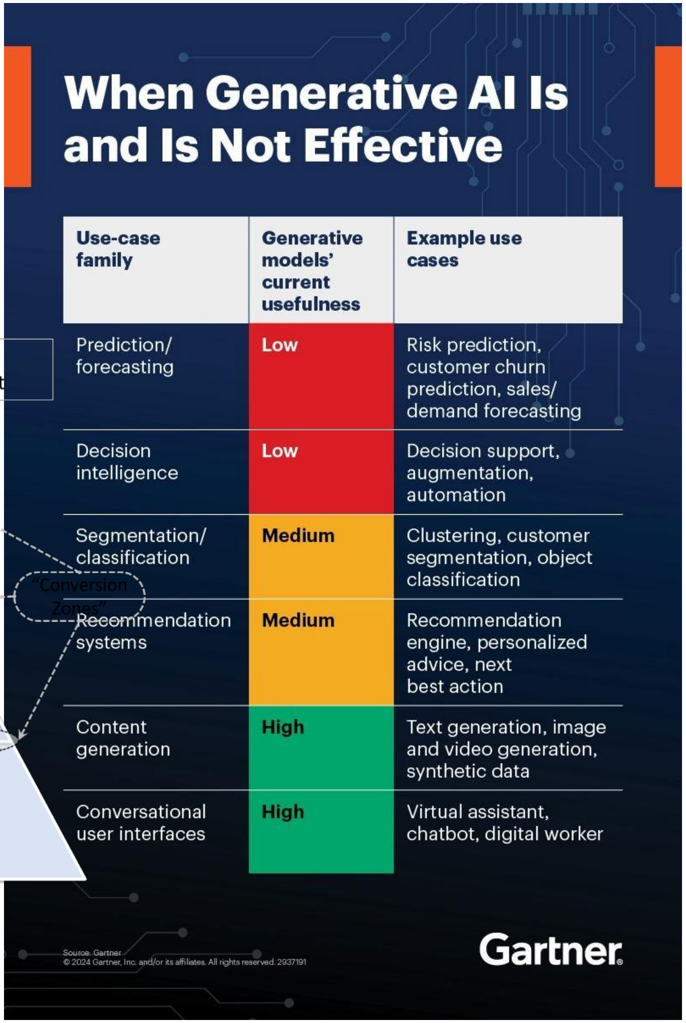
Singularity = Artificial General Intelligence



Machine / Human Consciousness



Source: Alphacution



When Generative AI Is and Is Not Effective

Use-case family	Generative models' current usefulness	Example use cases
Prediction/ forecasting	Low	Risk prediction, customer churn prediction, sales/ demand forecasting
Decision intelligence	Low	Decision support, augmentation, automation
Segmentation/ classification	Medium	Clustering, customer segmentation, object classification
Recommendation systems	Medium	Recommendation engine, personalized advice, next best action
Content generation	High	Text generation, image and video generation, synthetic data
Conversational user interfaces	High	Virtual assistant, chatbot, digital worker

Source: Gartner
© 2024 Gartner, Inc. and/or its affiliates. All rights reserved. 2937191

Gartner.

AGI? The Personalisation Trap!

‘Explainable’ AI identifies a new class of antibiotics (nature.com)

Mental health support for NHS patients with chatbot(s): Limbic | Clinical AI for Mental Healthcare Providers

Reference: Habicht, J., Viswanathan, S., Carrington, B. et al. (2024) Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nat Med 30, 595–602 https://doi.org/10.1038/s41591-023-02766-x

Overview ‹ Future You: Explore Your Future Self with Personalized Generative AI — MIT Media Lab

Deep Fakes for GriefBots:
Call for safeguards to prevent unwanted ‘hauntings’ by AI chatbots of dead loved ones | University of Cambridge

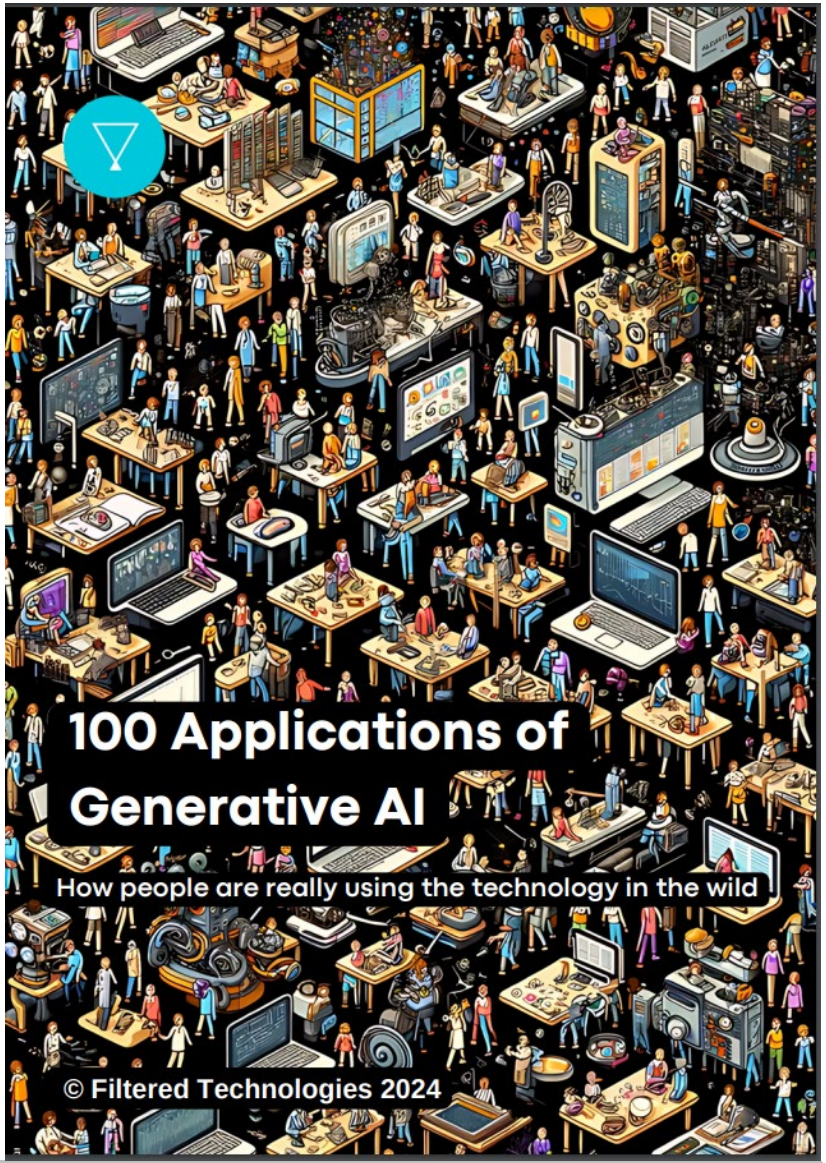
MIT takes down 80 Million Tiny Images data set due to racist and offensive content | VentureBeat

Reference: A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?," 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2021, pp. 1536-1546, doi: 10.1109/WACV48630.2021.00158.

GenAI? The Industry 5.0 Opportunity

- ✓ Content Creation and Editing
- ✓ Therapy / companionship
- ✓ Specific Search
- ✓ Explore topics of interest
- ✓ Creativity and recreation
- ✓ Troubleshoot
- ✓ Enhance learning
- ✓ Personalise learning
- ✓ Draft/ Adjust tone of email
- ✓ Simple explainers
- ✓ Draft/Summarise documents
- ✓ Edit CV
- ✓ Excel formulae
- ✓ Enhance decision-making
- ✓ Language translation
- ✓ Improve code
- ✓ Make complaints
- ✓ Cook with what you have

…



100 Applications of Generative AI

How people are really using the technology in the wild

© Filtered Technologies 2024

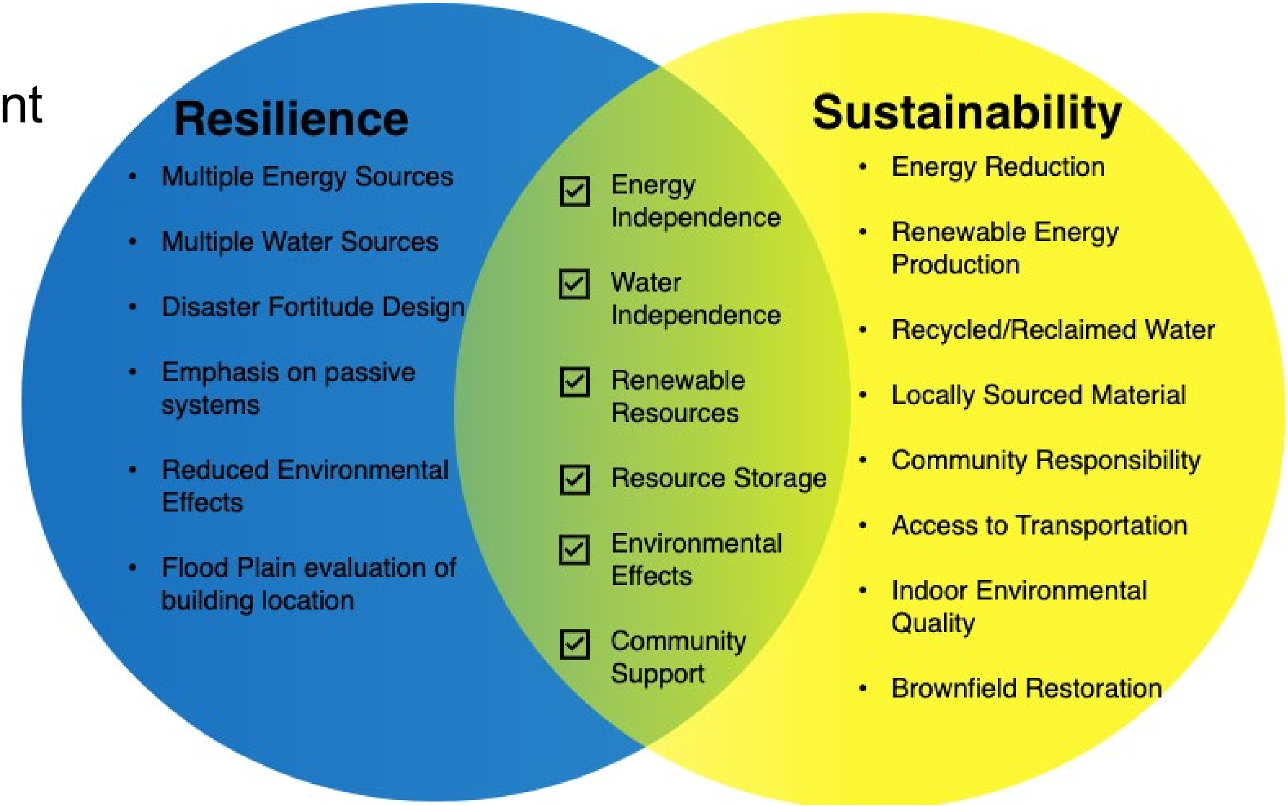
What is Sustainability?

Reference: Avery Sherffius and Smita Chandra Thomas How is Resilience related to Sustainability, Mitigation, and Adaptation? - energy-shrink.com

Sustainability and Resilience:

UN's 1987 World Commission on Environment and Development defines **SUSTAINABILITY** as "to meet the needs of the present without compromising the ability of future generations to meet their own needs".

Oxford dictionary defines **RESILIENCE** as "the capacity to recover quickly from difficulties".



Resilience

- Multiple Energy Sources
- Multiple Water Sources
- Disaster Fortitude Design
- Emphasis on passive systems
- Reduced Environmental Effects
- Flood Plain evaluation of building location

☑ Energy Independence

☑ Water Independence

☑ Renewable Resources

☑ Resource Storage

☑ Environmental Effects

☑ Community Support

Sustainability

- Energy Reduction
- Renewable Energy Production
- Recycled/Reclaimed Water
- Locally Sourced Material
- Community Responsibility
- Access to Transportation
- Indoor Environmental Quality
- Brownfield Restoration

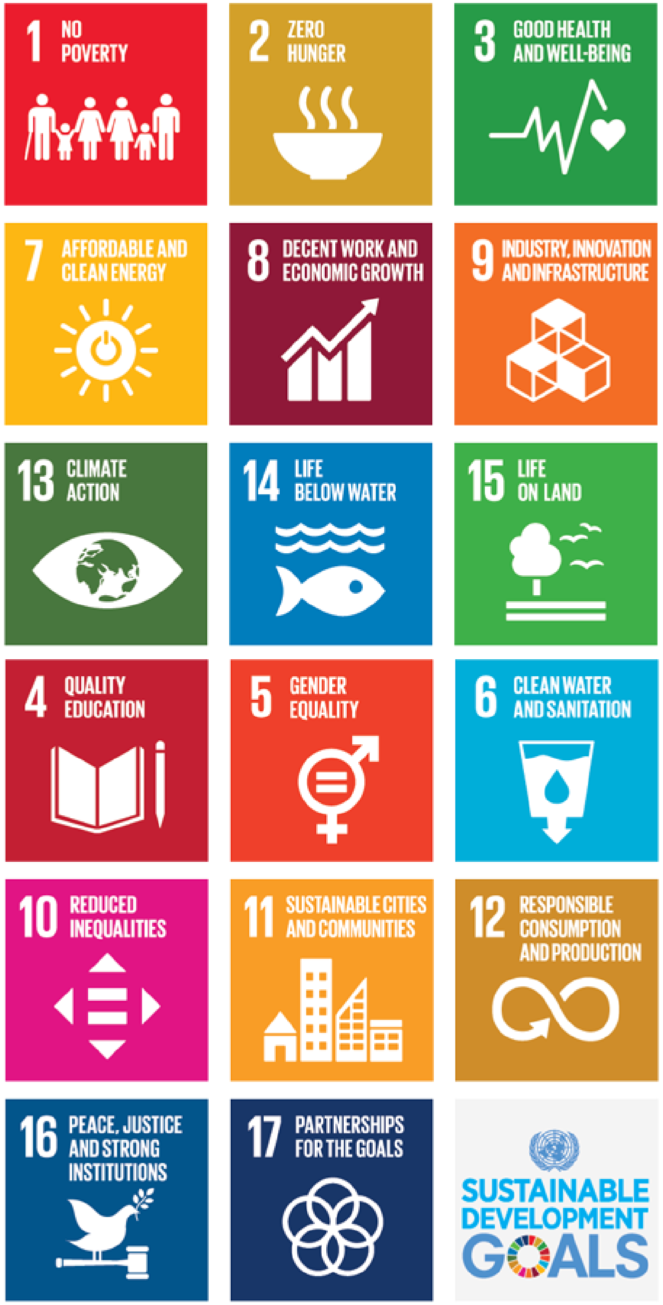
What are SDGs?

THE 17 GOALS | Sustainable Development (un.org), also known as the Global Goals, were adopted by the United Nations in 2015 as a universal call to action to end poverty, protect the planet, and ensure that by 2030 all people enjoy peace and prosperity.

The 17 SDGs are integrated—they recognize that action in one area will affect outcomes in others, and that development must balance social, economic and environmental sustainability.

Countries have committed to prioritize progress for those who're furthest behind. The SDGs are designed to end poverty, hunger, AIDS, and discrimination against women and girls.

The creativity, knowhow, technology and financial resources from all of society is necessary to achieve the SDGs in every context.


1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

What is Sustainable AI (SAI)?

SAI is the duty, obligation or expectation from technology (designers, owners or users) to provide or demonstrate features of:

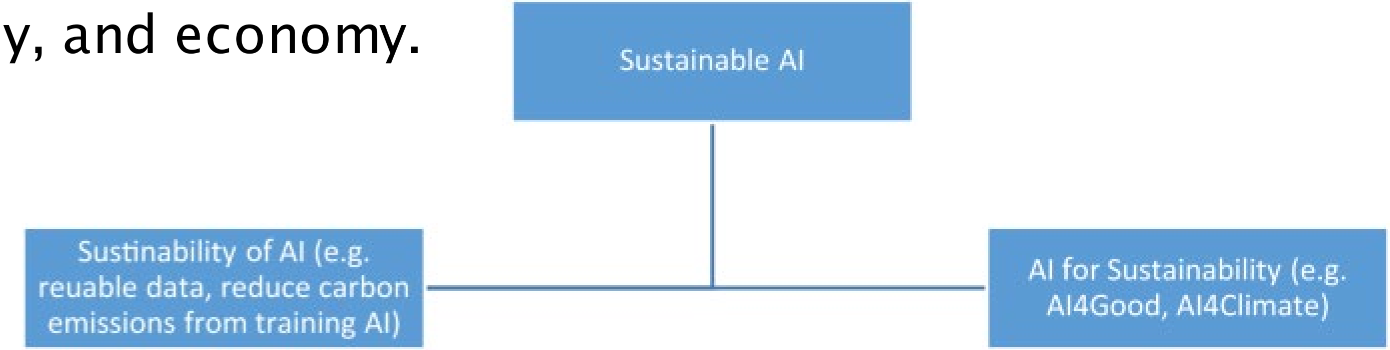
- GOVERNANCE;
- SUSTAINABLE AI: AI for sustainability and the sustainability of AI | AI and Ethics (springer.com)

Reference: van Wynsberghe, A. Sustainable AI: AI for sustainability and the sustainability of AI. AI Ethics 1, 213–218 (2021). https://doi.org/10.1007/s43681-021-00043-6

SAI is understood as having two branches:

AI for Sustainability + Sustainability of AI.

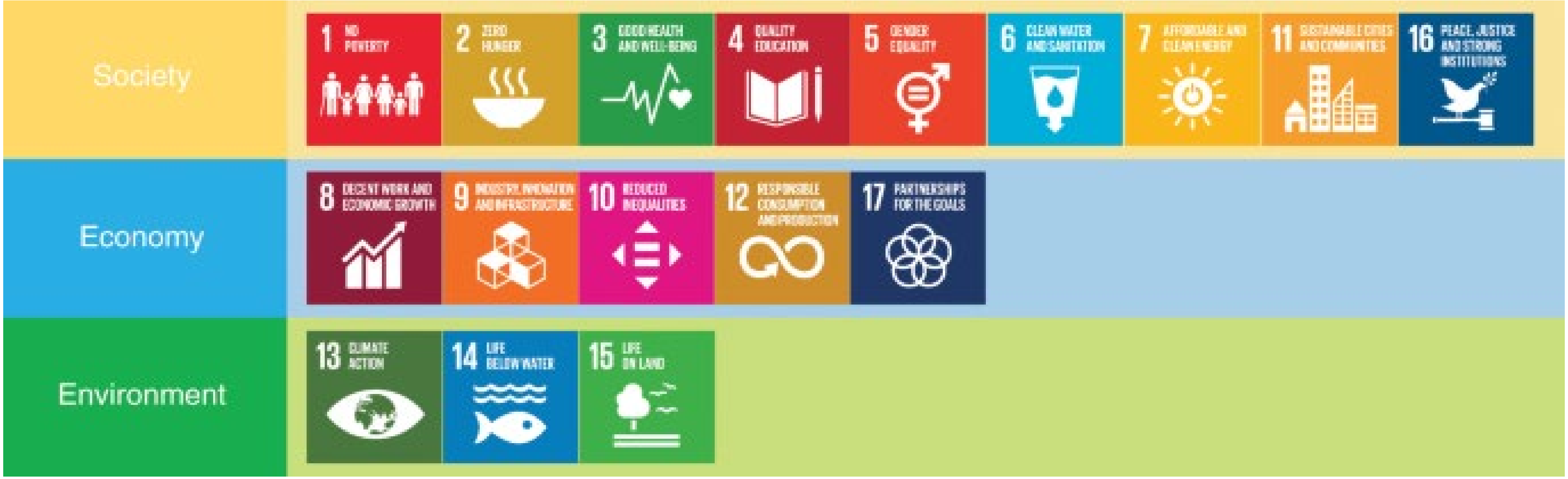
Sustainable AI takes sustainable development at the core of its definition with three accompanying tensions between AI innovation and equitable resource distribution; inter and intra-generational justice; and, between environment, society, and economy.


Sustainable AI
Sustainability of AI (e.g. reuable data, reduce carbon emissions from training AI)
AI for Sustainability (e.g. AI4Good, AI4Climate)

The role of AI in achieving SDGs?

The role of artificial intelligence in achieving the Sustainable Development Goals | Nature Communications:

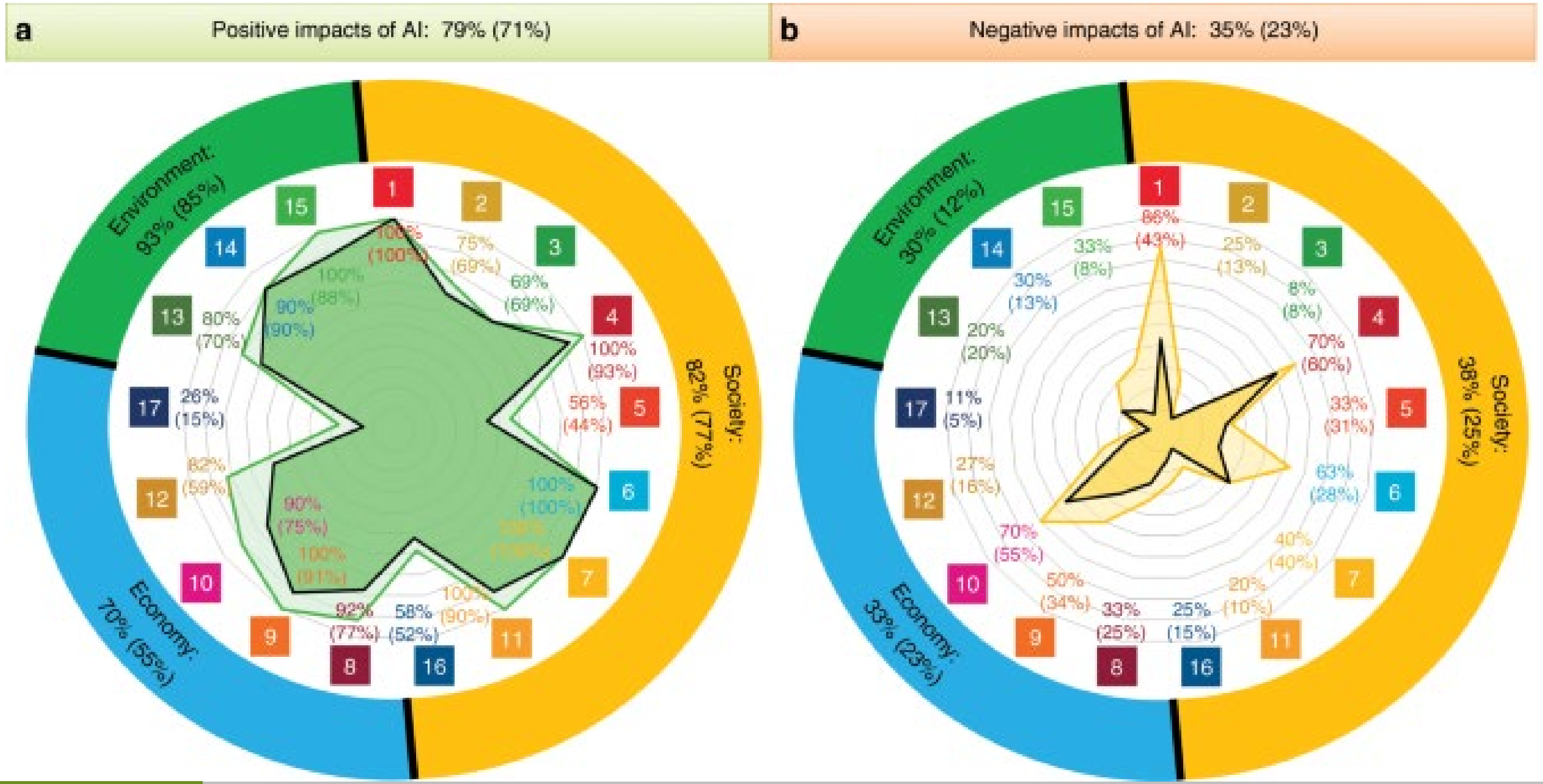
Reference: Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y


Society
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
11 SUSTAINABLE CITIES AND COMMUNITIES
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Economy
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
17 PARTNERSHIPS FOR THE GOALS
Environment
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND

The role of AI in achieving SDGs?

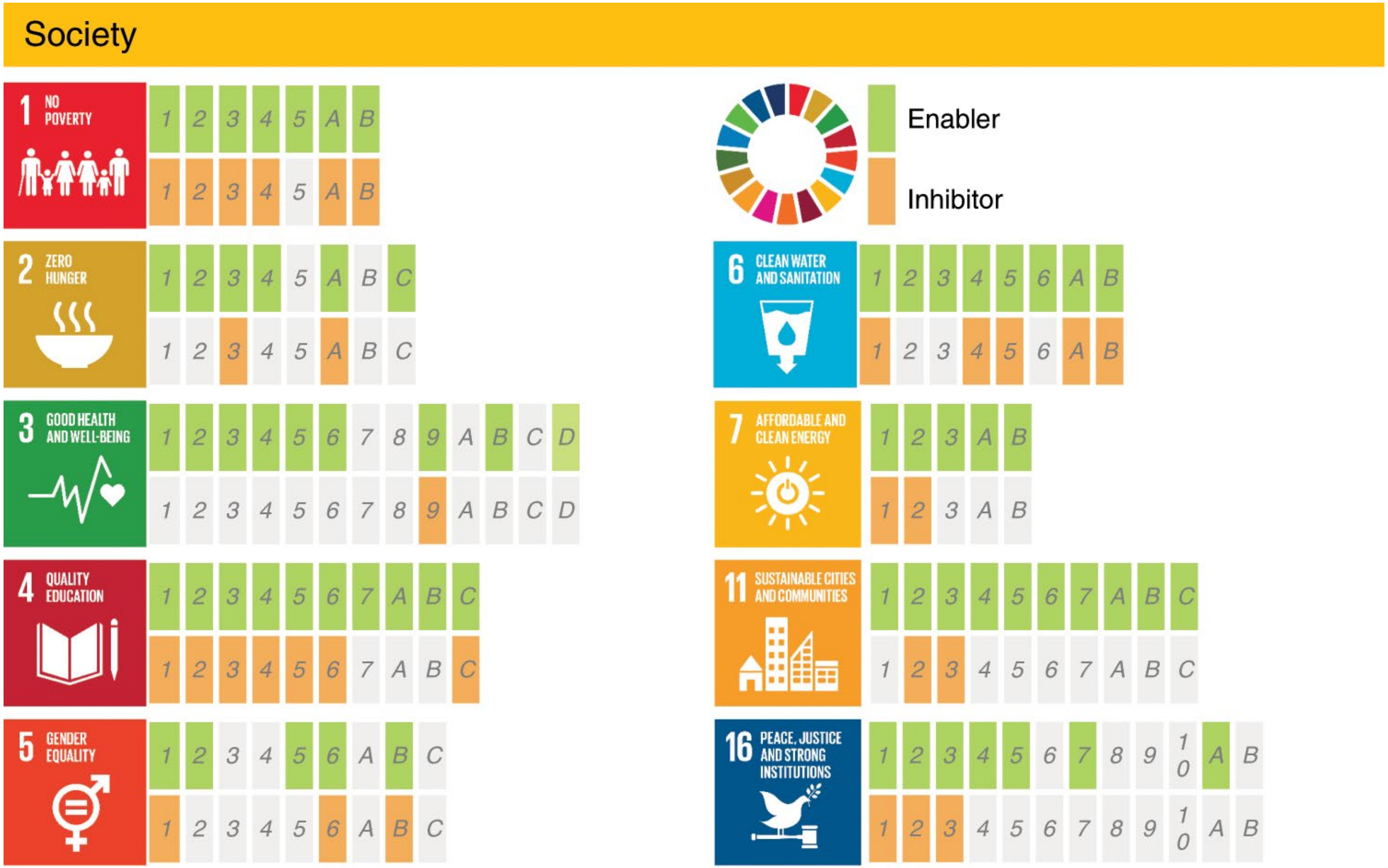
The role of artificial intelligence in achieving the Sustainable Development Goals | Nature Communications:

Reference: Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y



The role of AI in achieving WEHE SDGs?

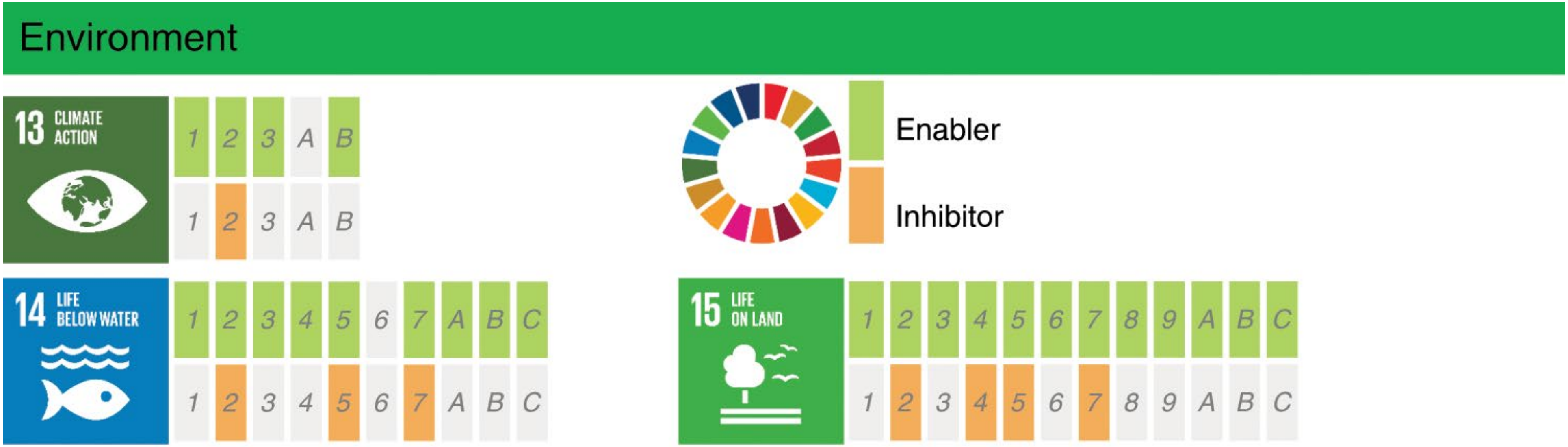
The role of AI in achieving WEHE SDGs | Nature Communications: 6, 7, 3, 13-15 – **the Society group**
Reference: Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y



The role of AI in achieving WEHE SDGs?

The role of artificial intelligence in achieving the Sustainable Development Goals | Nature Communications: the **Environment group**

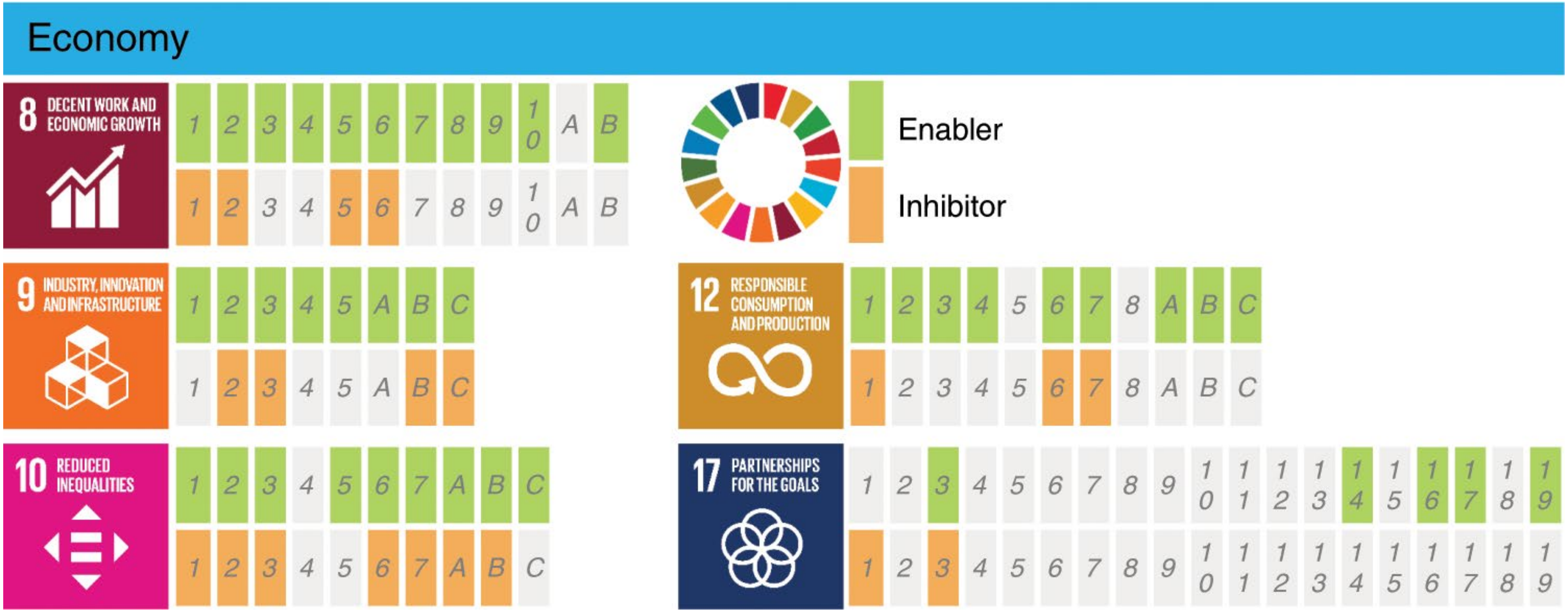
Reference: Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y



The role of AI in achieving WEHE SDGs?

The role of artificial intelligence in achieving the Sustainable Development Goals | Nature Communications: the **Economy group**

Reference: Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233 (2020). https://doi.org/10.1038/s41467-019-14108-y



Economy

🟩 Enabler

🟧 Inhibitor

8 DECENT WORK AND ECONOMIC GROWTH

	1	2	3	4	5	6	7	8	9	10	A	B
Enabler	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩
Inhibitor	🟧	🟧	⬜	⬜	🟧	🟧	⬜	⬜	⬜	⬜	⬜	⬜

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

	1	2	3	4	5	A	B	C
Enabler	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Inhibitor	⬜	🟧	🟧	⬜	⬜	⬜	🟧	🟧

10 REDUCED INEQUALITIES

	1	2	3	4	5	6	7	A	B	C
Enabler	🟩	🟩	🟩	⬜	🟩	🟩	🟩	🟩	🟩	🟩
Inhibitor	🟧	🟧	🟧	⬜	⬜	🟧	🟧	🟧	🟧	⬜

12 RESPONSIBLE CONSUMPTION AND PRODUCTION

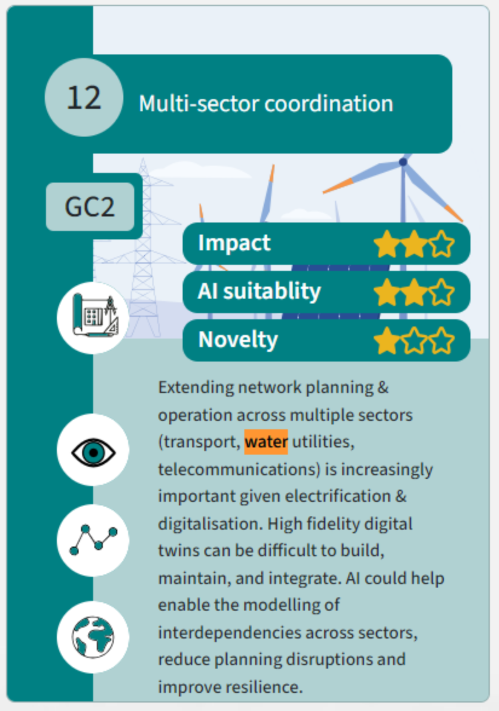
	1	2	3	4	5	6	7	8	A	B	C
Enabler	🟩	🟩	🟩	🟩	⬜	🟩	🟩	⬜	🟩	🟩	🟩
Inhibitor	🟧	⬜	⬜	⬜	⬜	🟧	🟧	⬜	⬜	⬜	⬜

17 PARTNERSHIPS FOR THE GOALS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Enabler	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟩	⬜	🟩	🟩	⬜	🟩
Inhibitor	🟧	⬜	🟧	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

The role of AI in achieving WEHE SDGs?

AI Decarbonisation Challenges - ADViCE | AI for Decarbonisation's Virtual Centre of Excellence (es-catapult.github.io)

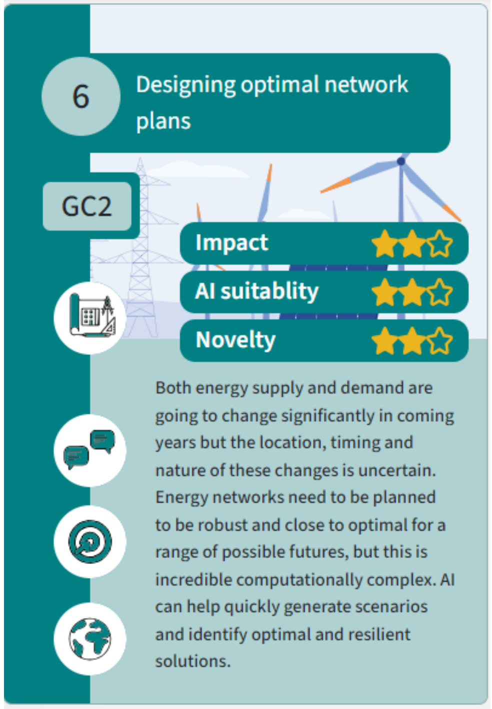


12 Multi-sector coordination

GC2

Impact	★★☆
AI suitablity	★★☆
Novelty	★☆☆

Extending network planning & operation across multiple sectors (transport, water utilities, telecommunications) is increasingly important given electrification & digitalisation. High fidelity digital twins can be difficult to build, maintain, and integrate. AI could help enable the modelling of interdependencies across sectors, reduce planning disruptions and improve resilience.

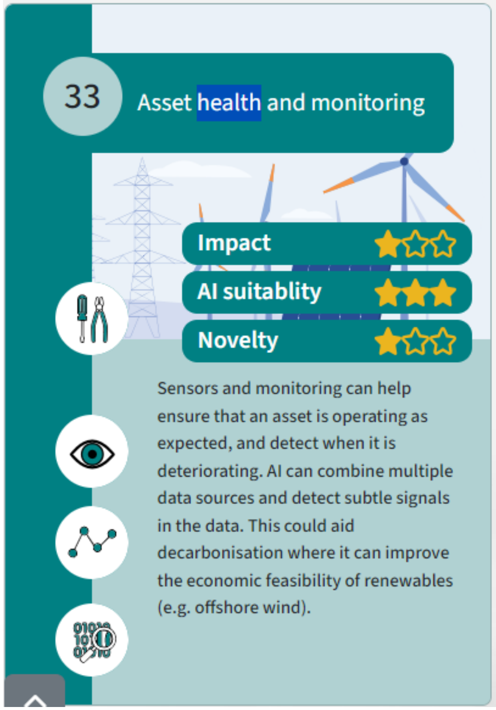


6 Designing optimal network plans

GC2

Impact	★★☆
AI suitablity	★★☆
Novelty	★★☆

Both energy supply and demand are going to change significantly in coming years but the location, timing and nature of these changes is uncertain. Energy networks need to be planned to be robust and close to optimal for a range of possible futures, but this is incredible computationally complex. AI can help quickly generate scenarios and identify optimal and resilient solutions.

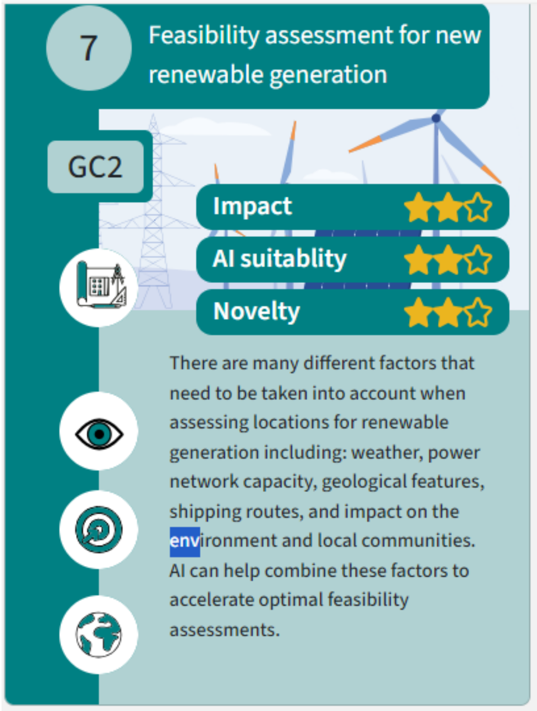


33 Asset health and monitoring

GC2

Impact	★☆☆
AI suitablity	★★★
Novelty	★☆☆

Sensors and monitoring can help ensure that an asset is operating as expected, and detect when it is deteriorating. AI can combine multiple data sources and detect subtle signals in the data. This could aid decarbonisation where it can improve the economic feasibility of renewables (e.g. offshore wind).



7 Feasibility assessment for new renewable generation

GC2

Impact	★★☆
AI suitablity	★★☆
Novelty	★★☆

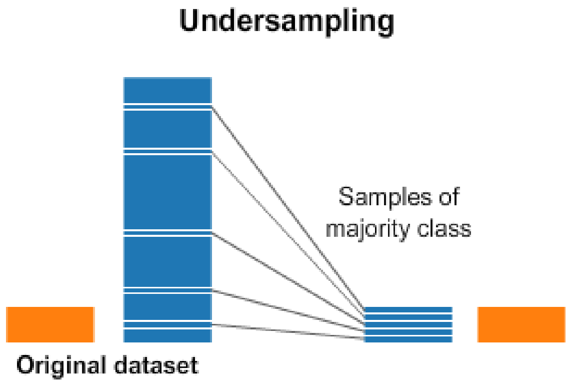
There are many different factors that need to be taken into account when assessing locations for renewable generation including: weather, power network capacity, geological features, shipping routes, and impact on the environment and local communities. AI can help combine these factors to accelerate optimal feasibility assessments.

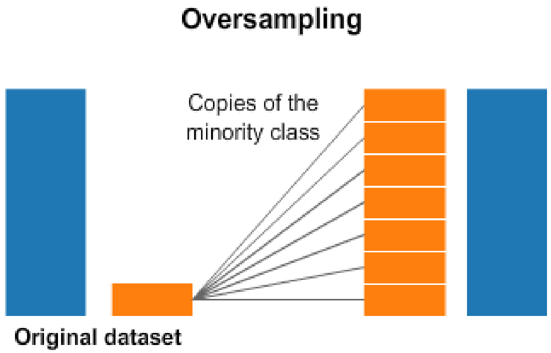
Do we need Sustainable Responsible AI?

Where are the AI risks and challenges generated from?

AI is shifting from Human Expert Knowledge to Machine Learning Models:

- **Quality and Relevant Data** to Any (publicly) available Big Data due to digital resource availability and business expectations, **using any imbalanced, biased, low quality, irrelevant training data**
- Statistical Learning grounds to Machine Learning automated solutions replacing Result Confidence with Model Output Performance and Accuracy
- Replacing Validation with Testing
- Expert Systems Industry Revolution to (Big) Data-driven/ -centric/ -enabled/ - enhanced AI models
- Lack of (Big) Data and AI Models Governance sustainable standards
- Decision support with robust models is replaced with Governing topics through numbers





Artificial GI? Natural GI? HGI?

Boston robot fights against pushing - BBC News

If we would like to describe what do we feel when we watch this "reinforcement learning" with supervised training and testing strategies, what would be the word you will use to label it?

AI technologies should assist us in our journey to understand, nurture, celebrate and sustain natural intelligence, wisdom and conscience: if this will be an individual and natural effort, a hybrid or society effort – we will see!

We shall start with evidence that **we ourselves** have:
- the complex and multifaceted ability to: learn (acquire; understand, process, retain) and apply knowledge; solve problems, reason, adapt to the environment, engage in abstract thinking and communications, *with*
- *Ethical, Transparent, Impartial (Just, Fair, Non-Maleficent), Reliable, Robust, Accountable, Trustworthy, Responsible, Wise and Conscientious manners.*

We can't predict your **future** but we can **help shape it**

Acknowledgment: undergraduate, postgraduate tutees, interns, PhD students and alumni, and academic colleagues with whom I collaborate on the SRAI topics