THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

AIAA 2290: Ethics, Privacy and Security in AI

Ethical Challenges and Responsibilities

Xuming HU

xuminghu@hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou)

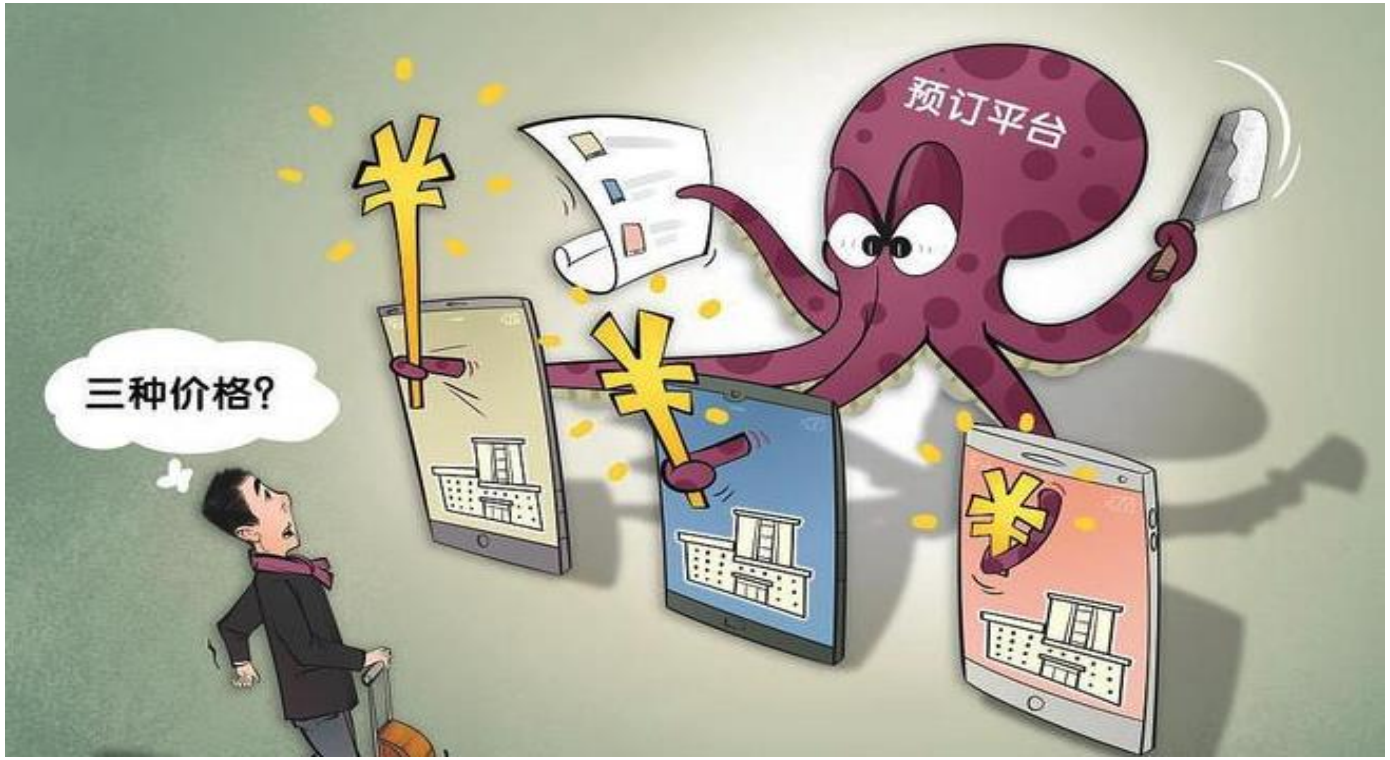
2025 Spring

Outline

What are the ethical challenges in AI?

Ethical Challenges in AI

- Some **e-commerce platforms** utilize **AI-driven recommendation** and pricing algorithms that adjust product or service prices based on *user data (e.g., browsing history, purchase frequency, device type, spending patterns).*
- Loyal or frequent shoppers may unwittingly be offered higher prices than new customers. This practice is often referred to as "**killing the familiar**," highlighting how the platform "exploits" repeat users' loyalty or habitual purchasing behavior.



Ethical Challenges in AI

AI systems are increasingly being used in healthcare for disease diagnosis and treatment recommendations. However, *AI models are not infallible, and diagnostic errors can occur due to biased training data, inadequate testing, or over-reliance on the system by medical professionals*. In some cases, such errors have led to misdiagnosis or treatment failures, potentially causing severe harm or even death to patients.

Ethical Challenges in AI

***Review last week's course content*:**

What principles were violated in the video?

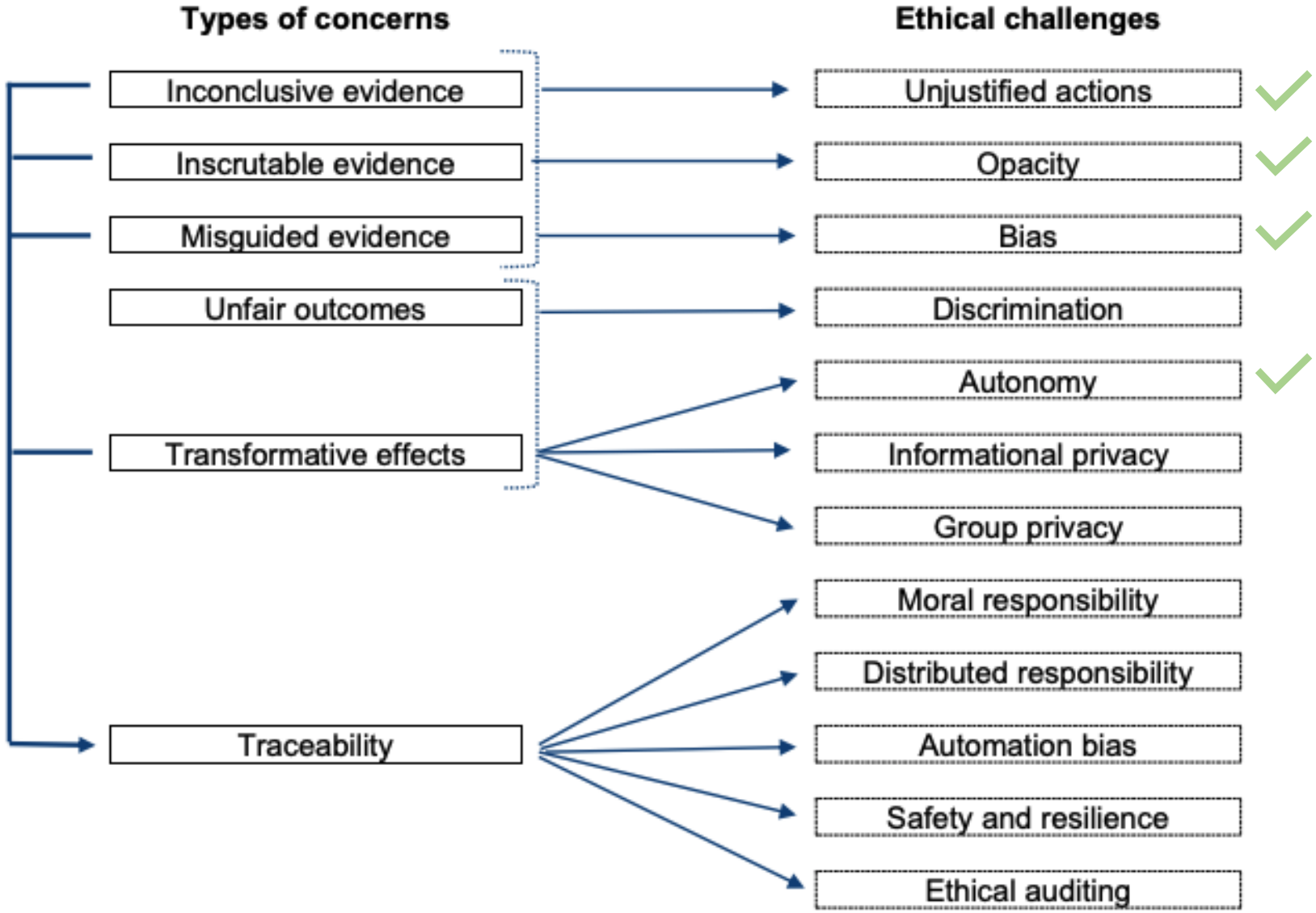
Ethical Challenges in AI

1. **Beneficence**
2. **Nonmaleficence**
3. **Autonomy**
4. **Justice**
5. **Explicability**

Ethical Challenges in AI

1. **Beneficence**
2. **Nonmaleficence**
3. **Autonomy**
4. **Justice**
5. **Explicability**

Ethical Challenges in AI



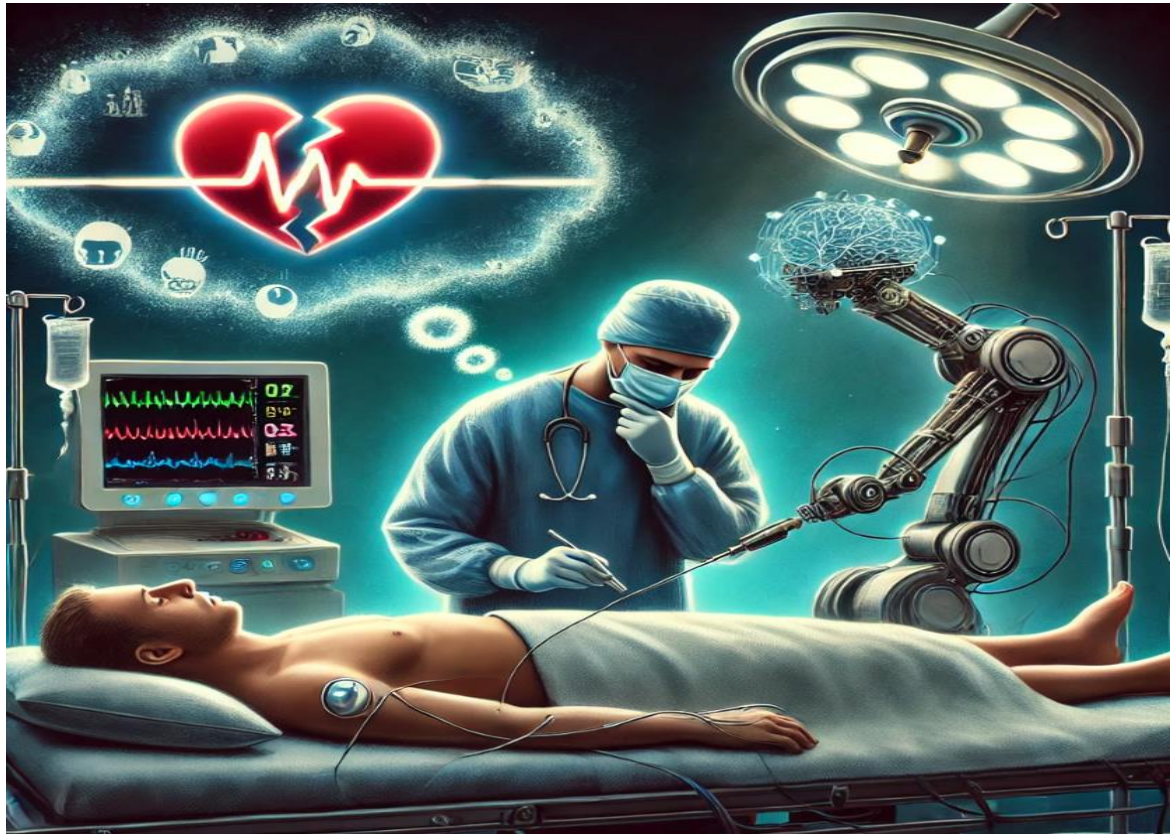
[1] Sauerbrei, A., Kerasidou, A., Lucivero, F., & Hallowell, N. (2023). The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Medical Informatics and Decision Making, 23*(1), 73.

Unjustified actions

Much algorithmic ***decision-making*** and ***data mining*** relies on inductive knowledge and correlations identified within a dataset. Correlations based on a 'sufficient' volume of data are often seen as sufficiently credible to direct action without first establishing causality. ***Acting on correlations*** can be doubly problematic. Spurious correlations may be discovered rather than genuine causal knowledge. Even if strong correlations or causal knowledge are found, this ***knowledge may only concern populations*** while actions with significant personal impact are directed towards individuals.

Unjustified actions

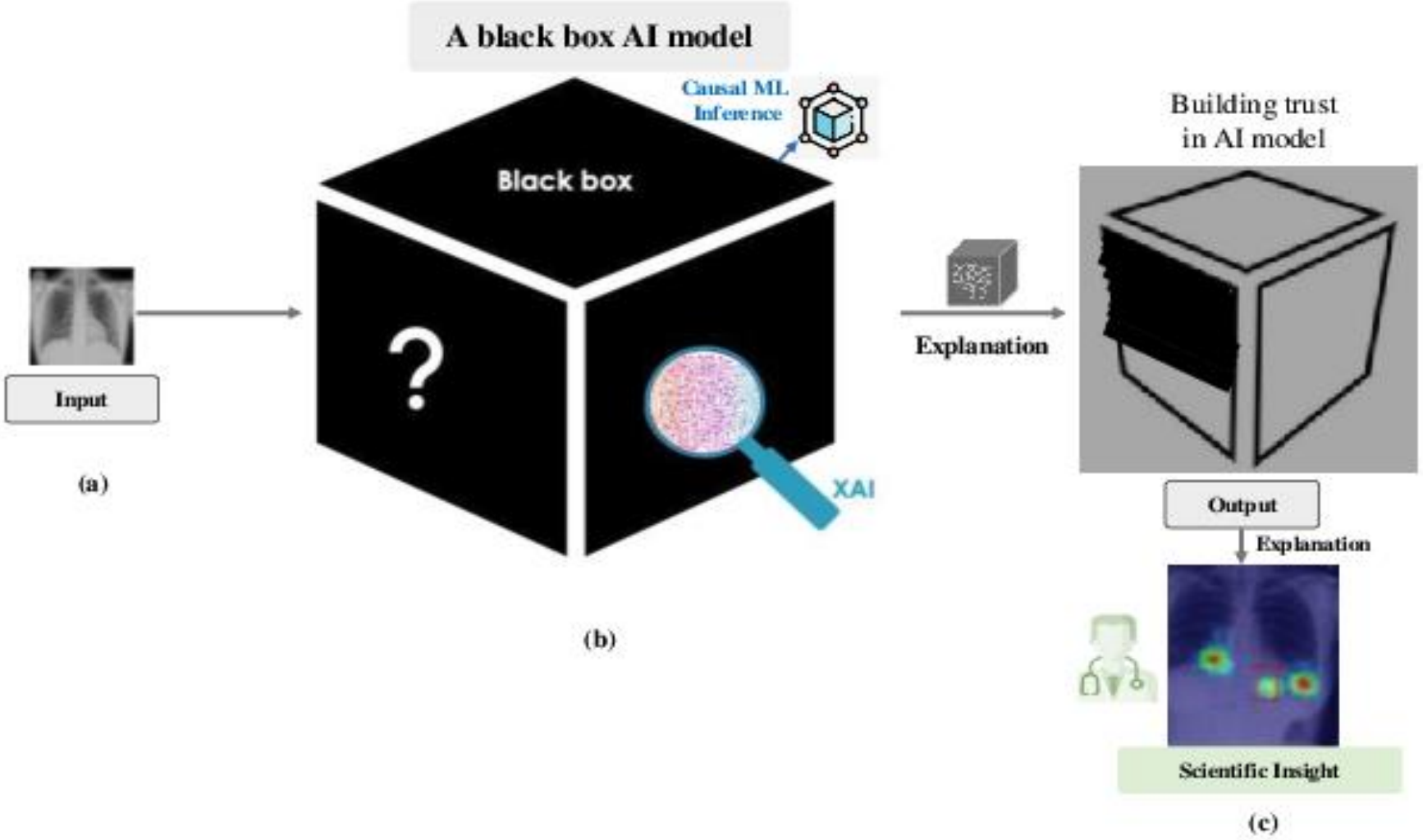
For example, algorithms designed to predict patient outcomes in clinical settings rely entirely on data inputs that can be quantified (e.g. vital signs and previous success rates of comparative treatments), whilst ignoring other emotional facts (e.g. the willingness to live) which can have a significant impact on patient outcomes, and thus, undermine the accuracy of the algorithmic prediction



Opacity

This is the ***'black box' problem*** with AI: the logic behind turning inputs into outputs may not be known to observers or affected parties or may be fundamentally inscrutable or unintelligible. Opacity in machine learning algorithms is a product of the high dimensionality of data, complex code and changeable decision-making logic. ***Transparency and comprehensibility*** are generally desired because algorithms that are poorly predictable or interpretable are difficult to control, monitor and correct.

Opacity


A black box AI model
Causal ML Inference
Black box
?
XAI
Input
(a)
(b)
Explanation
Building trust in AI model
Output
Explanation
Scientific Insight
(c)

Bias

The automation of human decision-making is often justified by an alleged lack of bias in AI and algorithms. This belief is unsustainable; AI systems unavoidably make biased decisions. A system's design and functionality reflects the values of its designer and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option. Development is not a neutral, linear path. As a result, "the values of the author, wittingly or not, are frozen into the code, effectively institutionalizing those values."

Bias

Inclusiveness and equity in both the design and usage of AI are thus key to combatting implicit biases. Friedman and Nisenbaum clarify that bias arises from:

- Pre-existing social values found in the “social institutions, practices, and attitudes” from which the technology emerges.
- Technical constraints.
- Emergent aspects of a context of use.

Bias

An example of Bias:

Please close your eyes and picture a shoe!

Bias

Did anyone picture this?

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

This?

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

How about this?

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

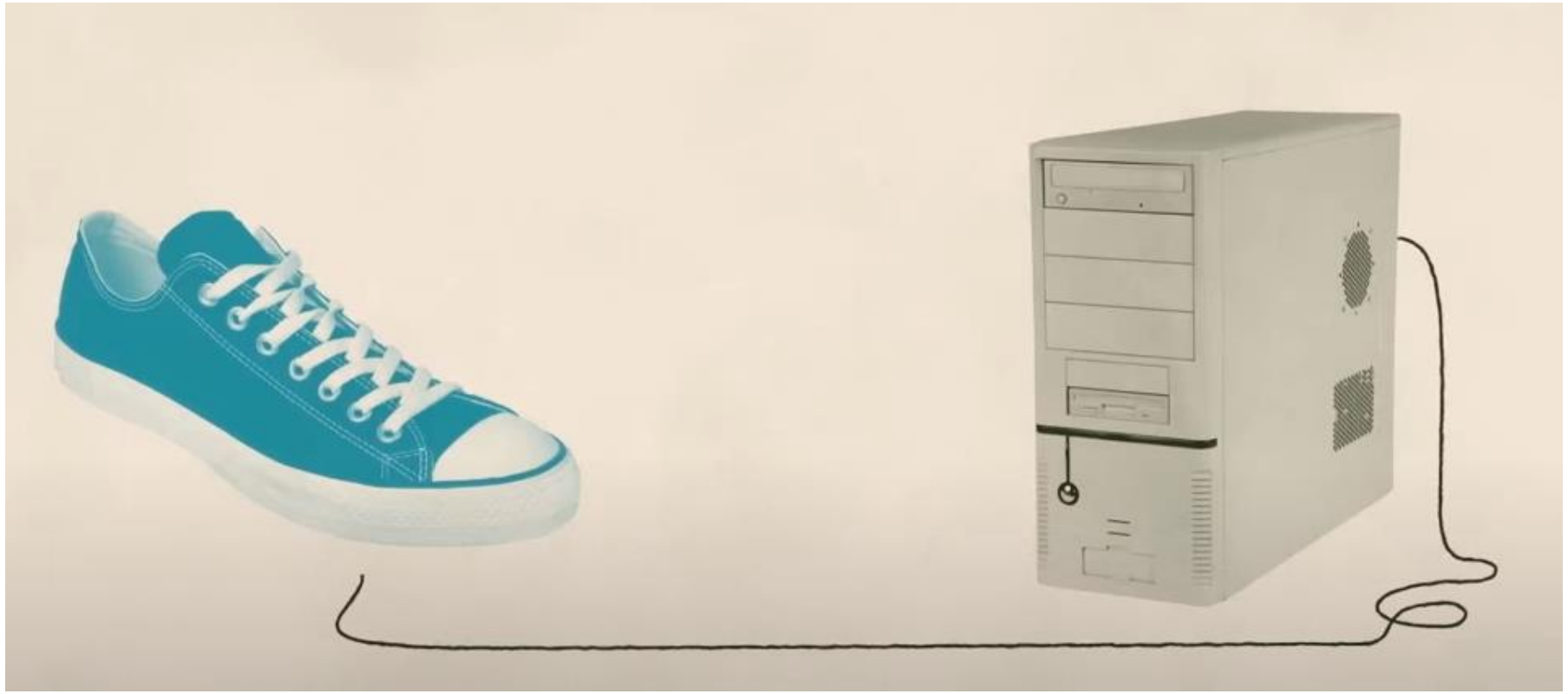
Bias

Each of us is biased toward one shoe over the others.

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

It is also biased for computer to recognize a shoe.



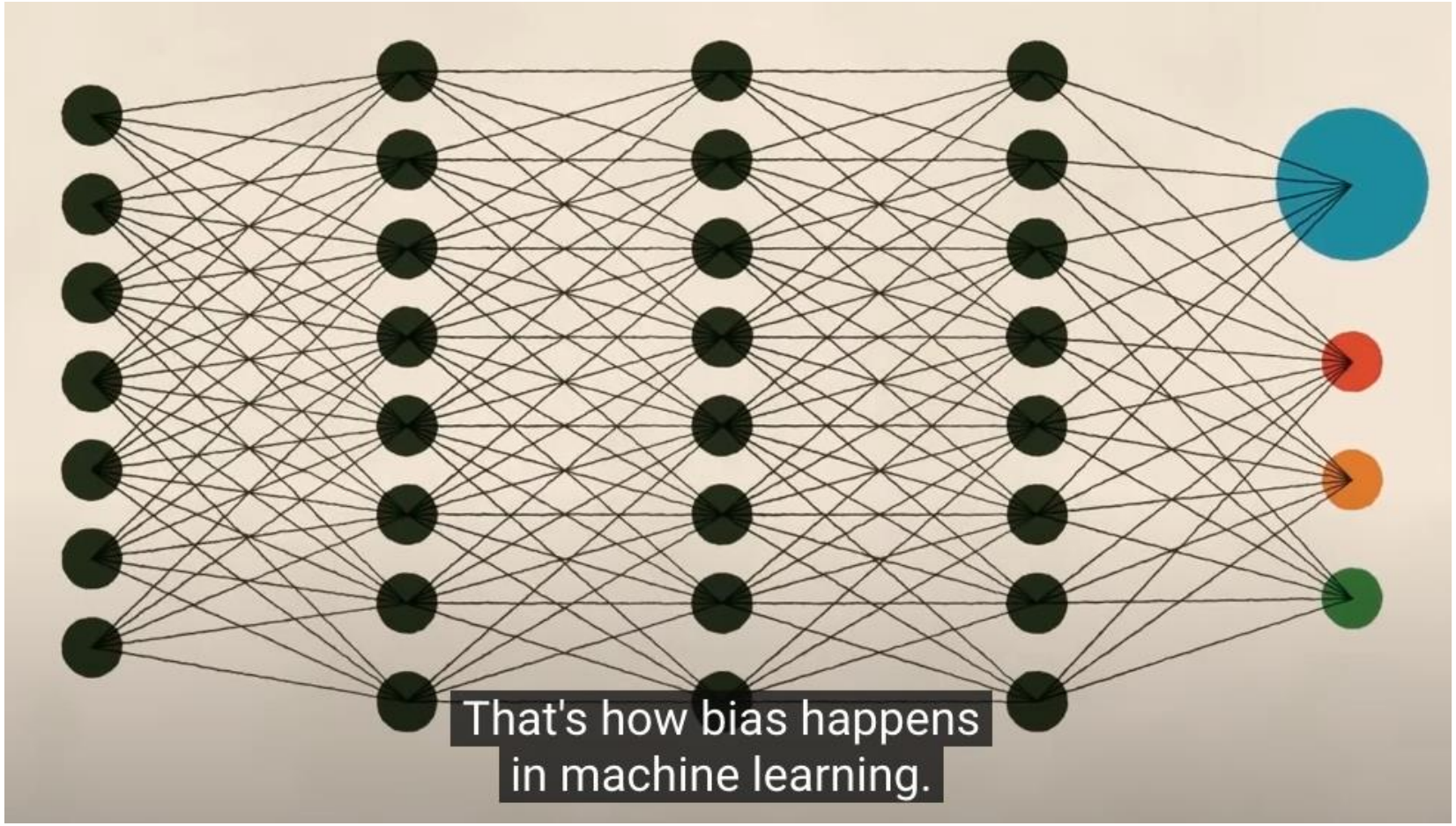
[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

It is also biased for computer to recognize a shoe.

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias



[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

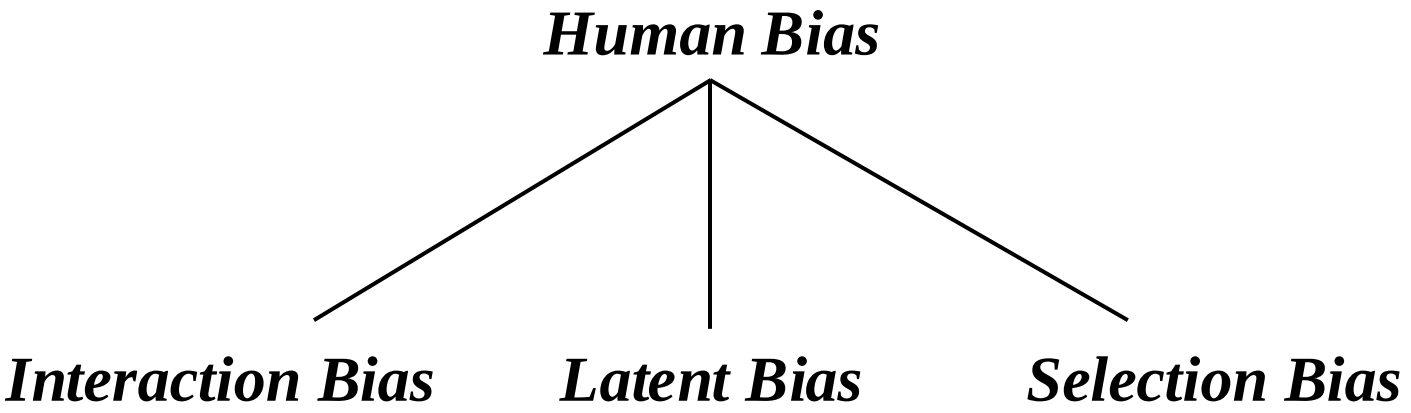
Data Bias

Tiny scale data

Large scale data

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

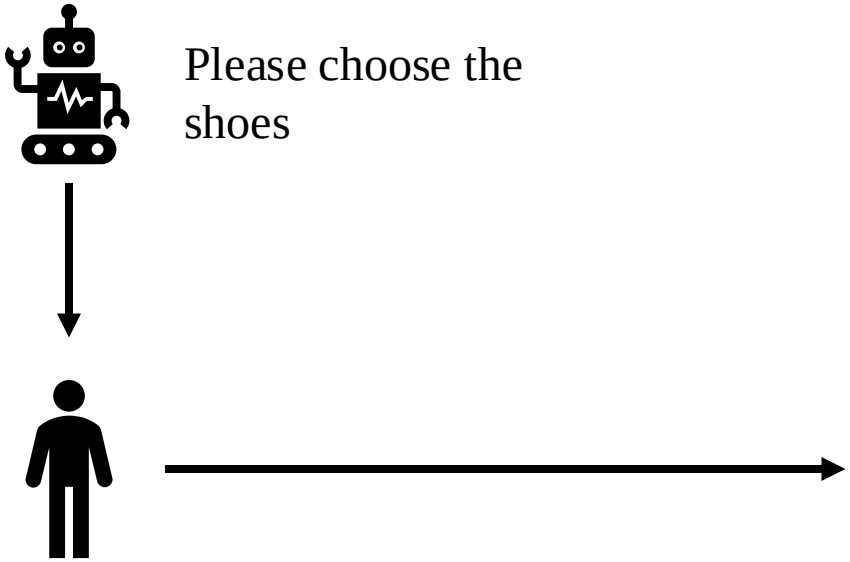
Bias

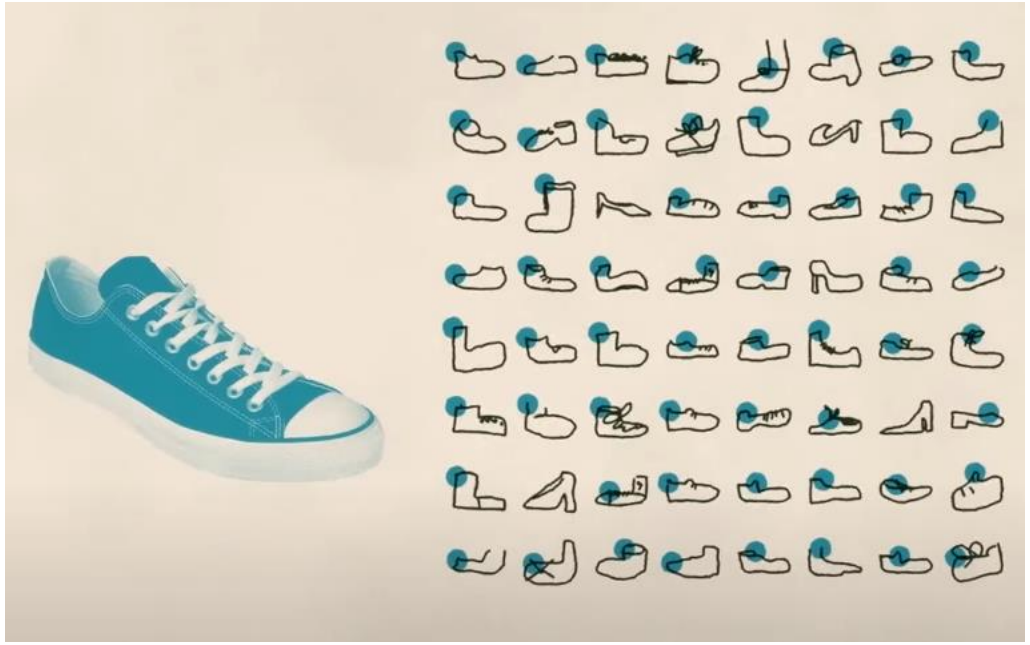


[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

Human Bias: Interaction Bias

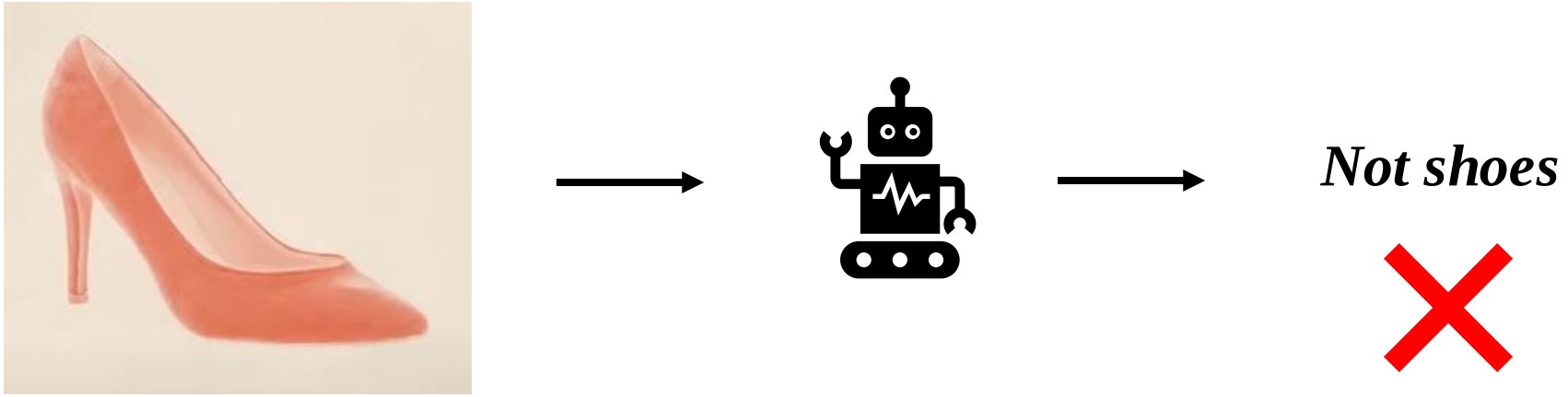




[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

Human Bias: Interaction Bias

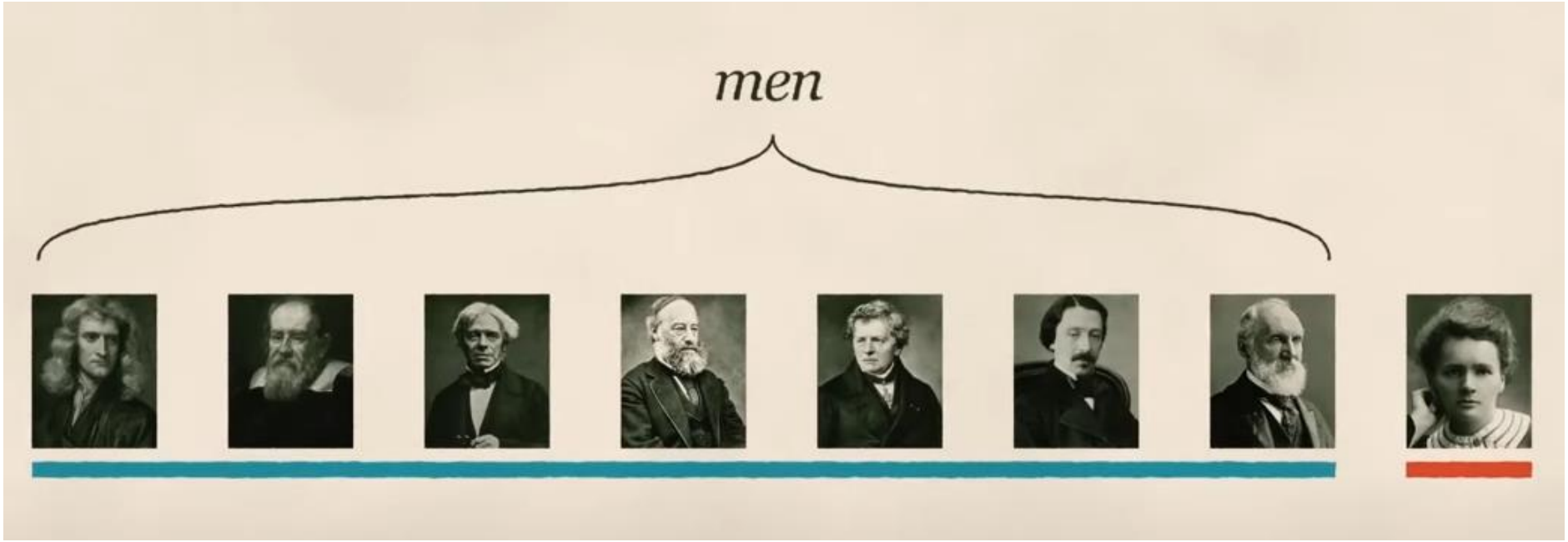


[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias

Human Bias: Latent Bias

When we train a model to recognize physicists



[1] https://www.youtube.com/watch?v=59bMh59JQDo.

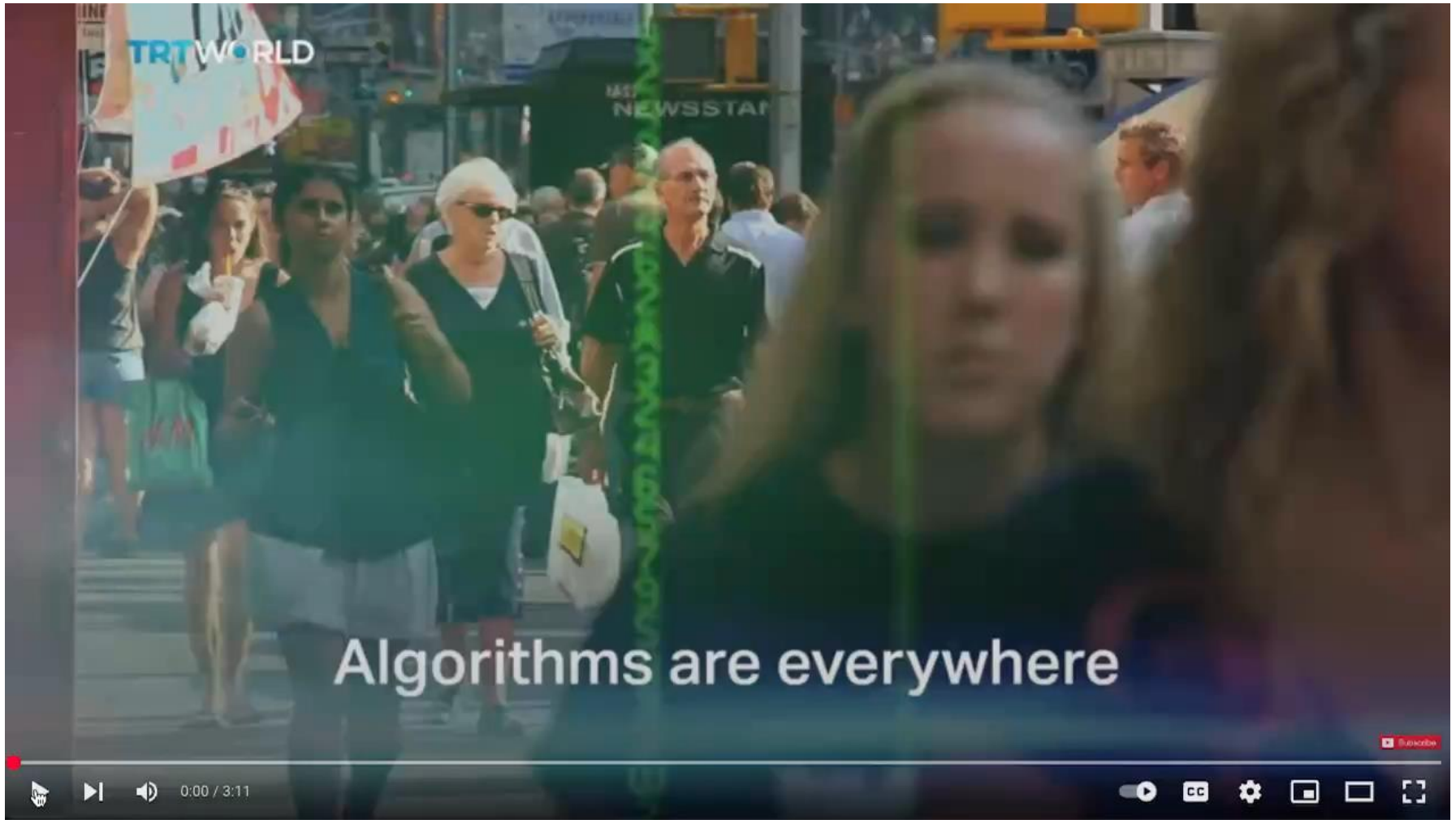
Bias

Human Bias: Latent Bias

When we train a face recognition model, the data is selected from a specific group of people

[1] https://www.youtube.com/watch?v=59bMh59JQDo.

Bias



[1] https://www.youtube.com/watch?v=bWOUw8omUVg.

Autonomy

Value-laden decisions made by algorithms can also pose a threat to ***autonomy***. Personalization of content by AI systems, such as recommender systems, is particularly challenging in this regard. Personalization can be understood as the construction of choice architectures which are not the same across a sample. AI can nudge the ***behavior of data subjects and human decision-makers*** by filtering information. Different information, prices, and other content can be offered to profiling groups or audiences within a population defined by one or more attributes.

Autonomy

For example: May experience a loss of autonomy if the basis for the recommendations is not well understood. Likewise, patients face a similar challenge when making informed decisions about their care based on AI recommendations. Recognising these risks, the WHO recognises “protecting human autonomy” as a key ethical principle for the design, usage, and governance of AI in healthcare due to the risk of decision-making power being transferred from humans to machines

Quiz of Ehical Challenges

In the context of AI used in facial recognition technology, which ethical challenges are implied?

A) Unjustified actions
B) Opacity
C) Bias
D) Autonomy

Quiz of Ehical Challenges

When implementing AI in hiring processes, which ethical challenges should be considered?

A) Unjustified actions
B) Opacity
C) Bias
D) Autonomy

Quiz of Ehical Challenges

When using AI in predictive policing, which ethical challenges could be present?

A) Unjustified actions
B) Opacity
C) Bias
D) Autonomy

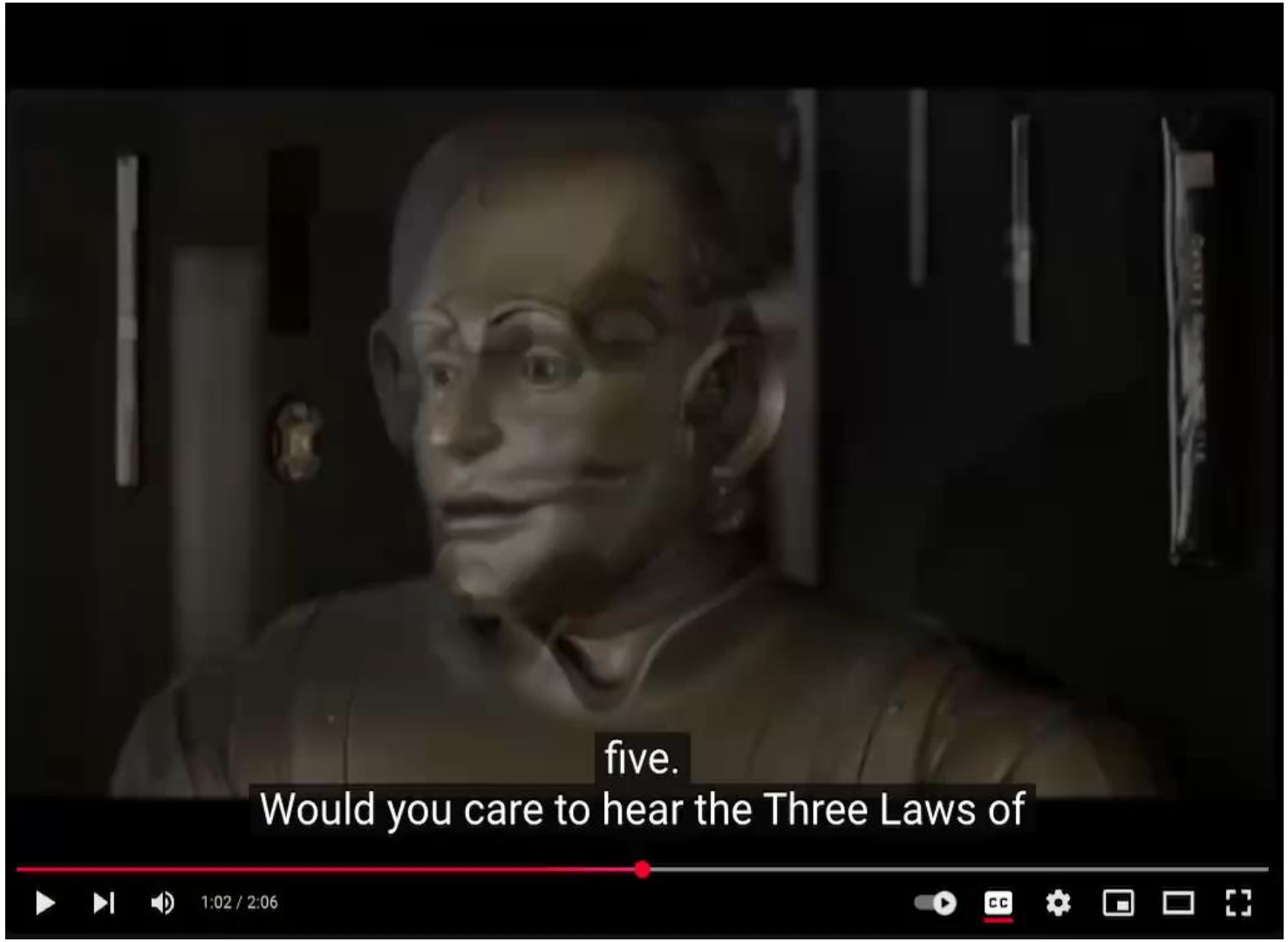
Quiz of Ehical Challenges

In the healthcare sector, particularly with AI diagnostic tools, which ethical challenges are relevant?

A) Unjustified actions
B) Opacity
C) Bias
D) Autonomy

Three Laws of Robotics

The Three Laws of Robotics ***(often shortened to The Three Laws or Asimov's Laws) are a set of rules devised by science fiction author Isaac Asimov, which were to be followed by robots in several of his stories.***



[1] https://www.youtube.com/watch?v=MsvdXY4lZag.

Three Laws of Robotics

First Law

A robot shall not harm a human being, or by inaction allow a human to be harmed.

Second Law

A robot shall obey any instructions and orders given to it by a human, except orders that would conflict with the First Law

Third Law

A robot shall avoid actions or situations that could cause to harm itself. Must protect its own existence if it not conflicts with the First and Second Law.

Three Laws of Robotics

[1] https://www.youtube.com/watch?v=958qOQ6LVuU&pp=ygUWdGhyZWUgbGF3cyBvZiByb2JvdGljcw%3D%3D.

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

- **Total Questions**: There are **8 questions** in this quiz. Each question is either True/False or Choice.
- **Participation**: Students compete to answer each question ***as quickly as possible***. The first student to raise their hand gets to answer.
- **Answering**:
  - If the ***first responder*** answers ***correctly***, they earn a point and get ***a small reward***.
  - If the answer is **incorrect**, another student gets the chance to answer.
- **Fair Play**:
  - Students must wait until the question is fully read before raising their hands to answer.
  - If a student interrupts before the question is completed, they forfeit their chance to answer that question.
- **Timing**:
  - Each student has ***10 seconds*** to answer once called on. If they don't respond in time, the chance moves to the next participant.

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q1:

A robot must always obey a human's orders, even if the order causes harm to another human.

True or False

Answer: False

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q2:

If a robot is ordered to self-destruct by a human, it must follow the command as long as no humans are harmed.

True or False

Answer: True

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q3:

A robot can harm a human if doing so will save its own existence.

True or False

Answer: False

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q4:

A robot must protect itself as long as it doesn't conflict with its duty to protect humans or obey orders.

True or False

Answer: True

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q5:

A robot sees two humans in danger. It can only save one due to time constraints. What should it do according to the Three Laws?

A. Save neither, as both options will lead to harm.
B. Randomly pick one to save.
C. Prioritize saving the human in greater danger.

Choose one answer

Answer: C

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q6:

A robot is asked to withhold critical life-saving information from a human. What should the robot do?

A. Obey the command to withhold information.
B. Disregard the command and share the information to prevent harm.
C. Prioritize its own existence by staying neutral.

Choose one answer

Answer: B

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q7:

If a robot receives two conflicting commands from two humans, what should it prioritize?

A. Follow the command from the higher authority figure.
B. Follow the command that causes the least harm.
C. Ignore both commands to avoid conflict.

Choose one answer

Answer: B

Three Laws of Robotics

Quiz Rules: Based on the Three Laws of Robotics

Q8:

What happens if a robot's actions to protect itself unintentionally harm a human?

A. This violates the First Law.
B. This is acceptable under the Third Law.
C. This follows the Second Law.

Choose one answer

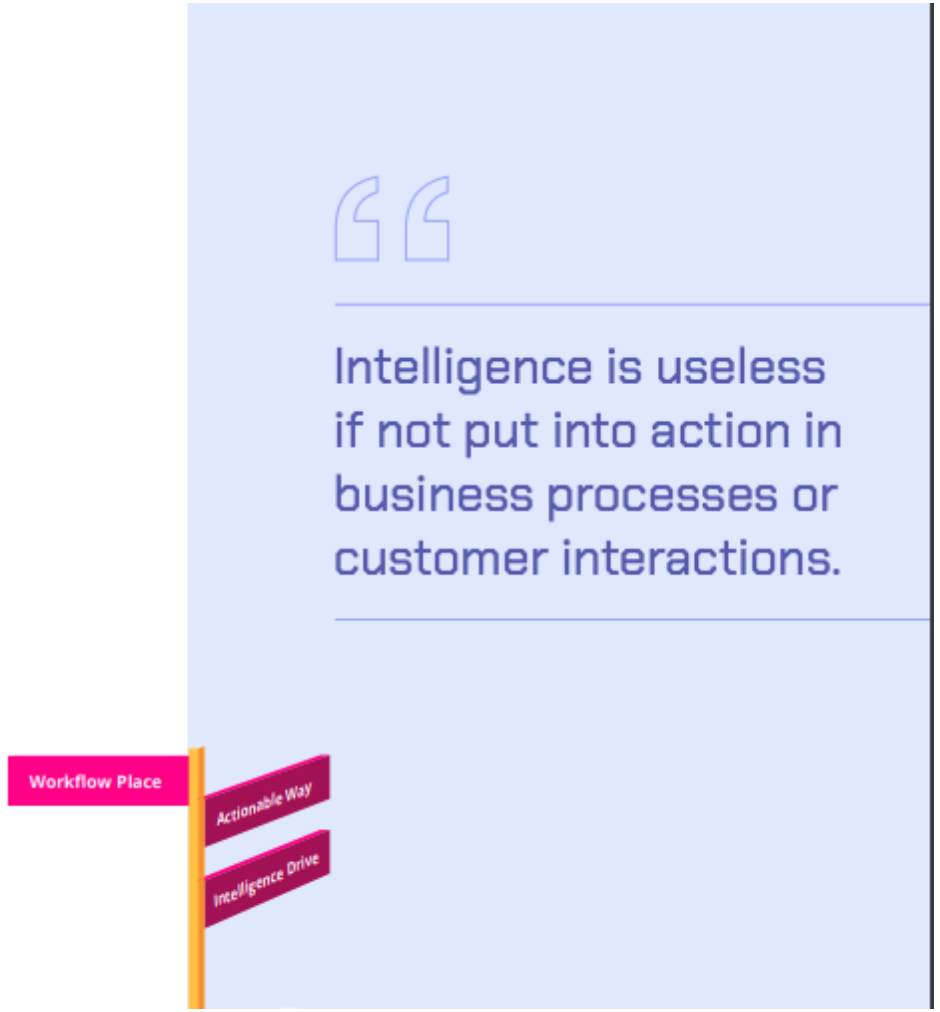
Answer: A

Nine Principles for Responsibility

01.

AI without action is a science experiment with no path to value

Intelligence is useless if not put into action in business processes or customer interactions. Start with the real business outcomes you want to achieve to understand which AI-powered decisions and generative models drive the most value. Then identify the specific workflows and customer interactions you want to impact. Having well-structured interactions, workflows, and case management creates the data and feedback loop required for additional optimization through artificial intelligence.



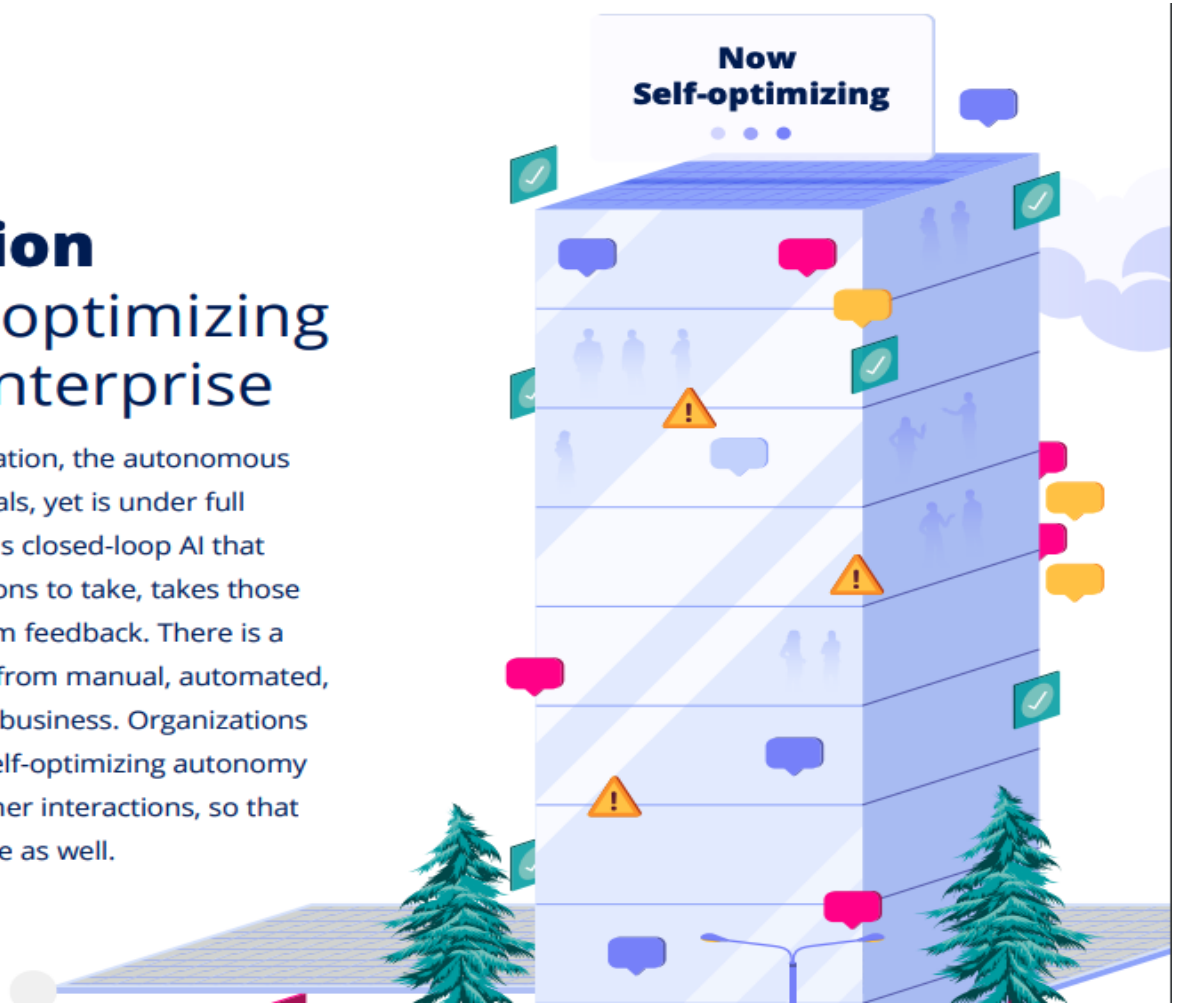
[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility

02.

AI & automation power the self-optimizing autonomous enterprise

Powered by intelligence and automation, the autonomous enterprise self-optimizes toward goals, yet is under full control of the business. This requires closed-loop AI that executes in real time, proposes actions to take, takes those actions, and immediately learns from feedback. There is a path toward autonomy that moves from manual, automated, intelligent business to autonomous business. Organizations should create a roadmap to build self-optimizing autonomy into their key workflows and customer interactions, so that AI becomes autonomous intelligence as well.



[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility

> ...AI is a tool designed by humans, and it is most valuable in instances where it is assisting...

AI STRATEGY

The what & how

03.

AI is augmented intelligence – **it's best with humans in control**

Hollywood movies and "AI doomer" blogs want to make you believe AI is all about taking control away from humans. The reality is that AI is a tool designed by humans, and it is most valuable in instances where it is assisting the agent, employee, developer, marketer, or customer by providing guidance and taking on tasks. To minimize risk, AI strategy should prioritize human control by considering "human-in-the-loop" use cases or ensuring humans are monitoring and steering autonomous AI in play.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility

04.

There is more to AI than just gen AI –
you need left & right brain AI

Generative AI is injecting creativity – right-brain thinking – into the way we do business and design applications. But many automated decisions in workflows and gen AI driven interactions require "left-brain AI," the analytical and rational AI capabilities that drive real-time automated decision-making, recommend the next best action across all channels, and proactively spot process inefficiencies before they become challenges. Generative AI may have gotten your C-suite excited about AI-powered transformation, but your AI strategy needs to cover all forms of AI to be complete, actionable, and successful.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility

05.

Start with outcomes & decisions, not with data and models

It is easy to get lost in data swamps, or marvel at all kinds of machine learning models that could be created. A solid and secure data strategy is important, but you must think top-down, not just bottom-up. What outcomes in the business do you want AI to optimize? What customer experiences do you want to improve? What are the automated decisions and generated intelligence that can drive these outcomes? Which models can automate the decisions? Answering these questions will help prioritize what data matters most to an organization.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility

06.

You'll need an open, **best-of-breed model** ecosystem

The debate between "public" and "private" AI is a distraction. Organizations will need to run private models – trained on their own data – for differentiated use cases such as customer engagement and process optimization. In other cases, publicly available models can provide a secure, fast, and scalable path to value for commoditized use cases, such as optical character recognition (OCR) and image recognition. Many generative AI foundation models and services are available to build upon. Establishing an AI architecture that protects private data while supporting private models, fine-tuning public and open-source models, and using public AI services provides the flexibility and security to adopt and scale the right AI, for the right use cases, in the timeframes markets will demand.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility

ETHICAL, RESPONSIBLE, & TRUSTWORTHY AI

The right way

07.

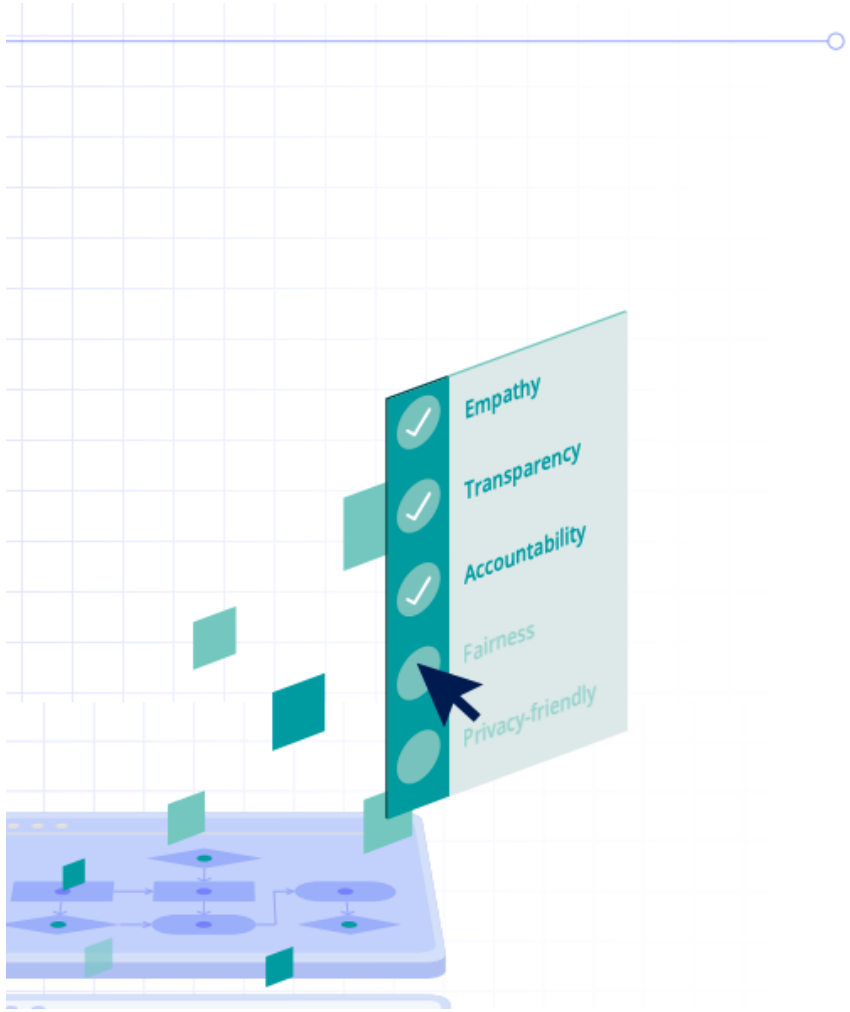
AI should be **empathic** to all stakeholders

Empathy is about putting yourself in others' shoes and doing what's right not just for you, but for everyone else. This is key for ensuring AI is trusted, responsible, accepted, wanted, and compliant. Employees and customers are fearful that organizations may use AI to drive profits at their expense. To ensure buy-in, your AI initiatives should deliver value to customers and employees, and striking the right balance should be ingrained at all levels of automated decisioning and AI.

To ensure buy-in, your AI initiatives should deliver value to customers and employees.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility



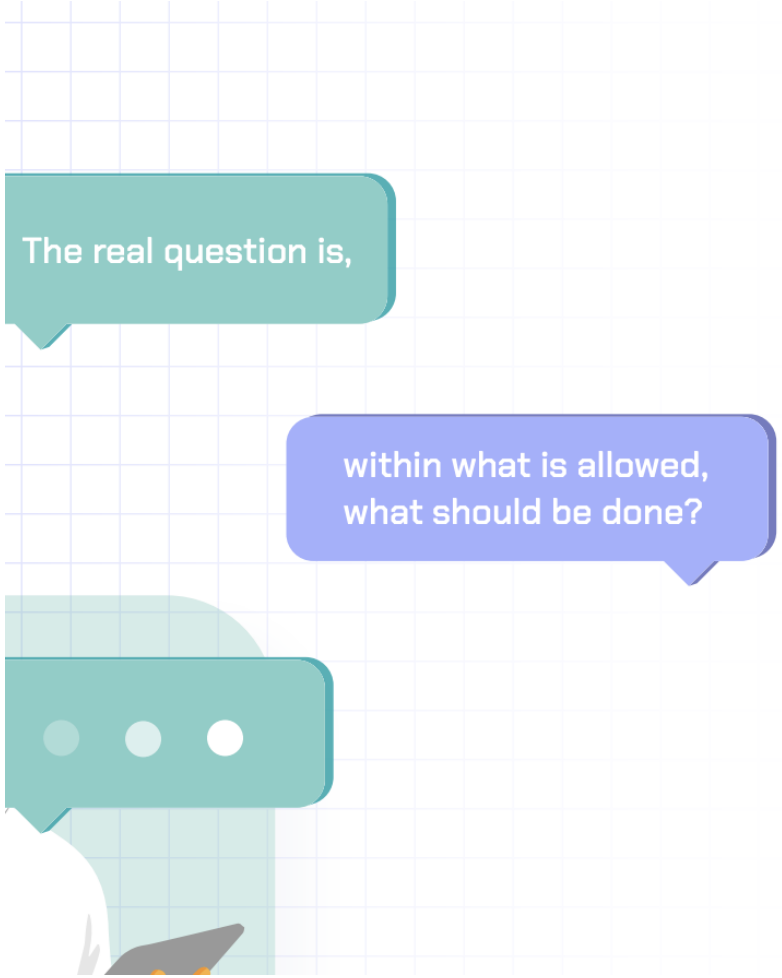
08.

Build ethical principles into your tools & processes

Everyone will state that AI needs to be fair, transparent, safe, accurate, robust, privacy-friendly, auditable, managed, and accountable – promoting organizational benefits as well as societal and environmental well-being. You must bake these principles into your tools and best practices. Understand and govern which decisions and use cases require explainable AI, and where more opaque models are acceptable. Explanations are not just for data scientists but should also target end users such as agents and customers. Create processes that continuously check for hidden bias in models and automated decisions. Ensure all automated decisions and gen AI interactions are audited – both for you and your customers.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

Nine Principles for Responsibility



09.

AI ethics goes beyond **regulatory compliance**

The U.S., EU, and other governments globally are moving rapidly to create guidelines and laws for the use of AI, but it's a trap to think about AI ethics and responsible AI solely in terms of what is allowed and what is not. The real question is, within what is allowed, what should be done? For instance, using AI to promote conversations that benefit customers and are relevant to their needs will deliver far greater payoff in the long run than pushing sales opportunities in the short term. Ensure that the customer is at the center of your AI strategy by driving outcomes that deliver both a great customer experience and improvements to your bottom line – turning an ethical and empathetic approach into a competitive advantage.

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

[1] https://www.youtube.com/watch?v=dWRnCXbUDgA

Six big ethical questions about the future of AI:

1	**Is the system autonomous?**
2	**Does the system have agency/ control?**
3	**How do we think about its assurance/ safety/ does it function properly?**
4	**Will there be interface between systems?**
5	**What will the indicators be that show that systems work well/ efficiently?**
6	**What is the systems' intent, what is it designed to do?**

[1] https://www.pega.com/the-ai-manifesto: pega-ai-manifesto-v3

There are some **loopholes** of the three laws. For example:

A robot doctor refuses to perform surgery that could save a life because the procedure has a 50% chance of failure, risking harm to the patient.

Propose New Laws:

- Suggest new or modified laws to **address the identified loophole**. Laws should aim to resolve the conflict without violating the fundamental principles of robotics.
- Use AI tools like **ChatGPT** to brainstorm ideas, analyze cases, or refine their proposed laws.
- Upload your proposal **on the Canvas (10 minites)**
- Anyone that actively shares the AI insights with the class earns bonus a small reward.
- ***Practice is not included in the final score***

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

AIAA 2290: Ethics, Privacy and Security in AI

Thanks!!

Xuming HU

xuminghu@hkust-gz.edu.cn

The Hong Kong University of Science and Technology (Guangzhou)

2025 Spring