

# Mapping the Minefield: Identifying and Assessing AI Risks for Internal Auditors

 by Antony Hibbert



# Tony Hibbert

AI Governance Expert

*Simplifying AI Governance*

antony\_hibbert@outlook.com

<https://www.linkedin.com/in/antonyhibbert-ai-governance-expert/>



**Antony Hibbert**

Experienced AI Governance Expert |  
Navigating the path to responsible AI for fin...



# Agenda

- Understanding AI and its integration
- Trends and evolving risks in AI
- Regulatory and ethical considerations
- Identifying AI risks and assessing AI risks in practice

# Introduction

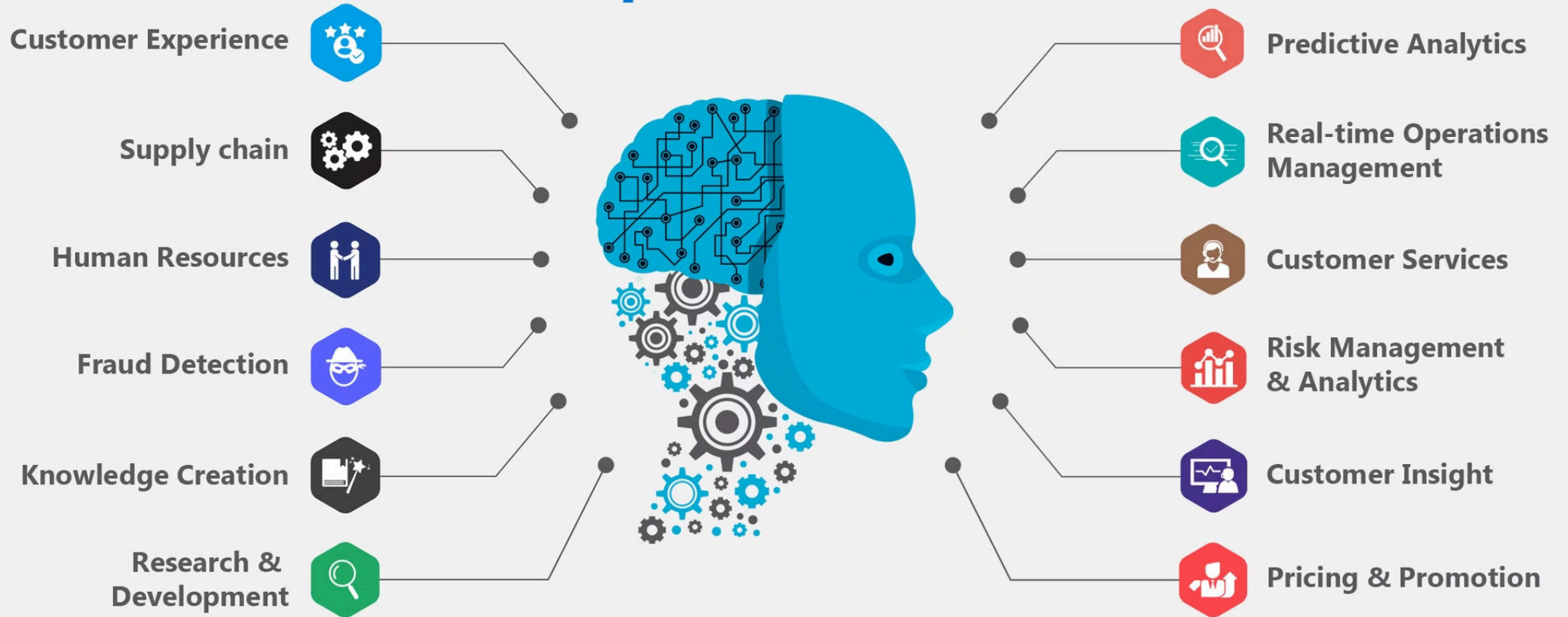
- Importance of AI in current audit environments
- Equipping internal auditors with the knowledge to navigate AI risks



# Understanding AI and its Integration

# Some Use cases of AI

## Top AI Use Cases

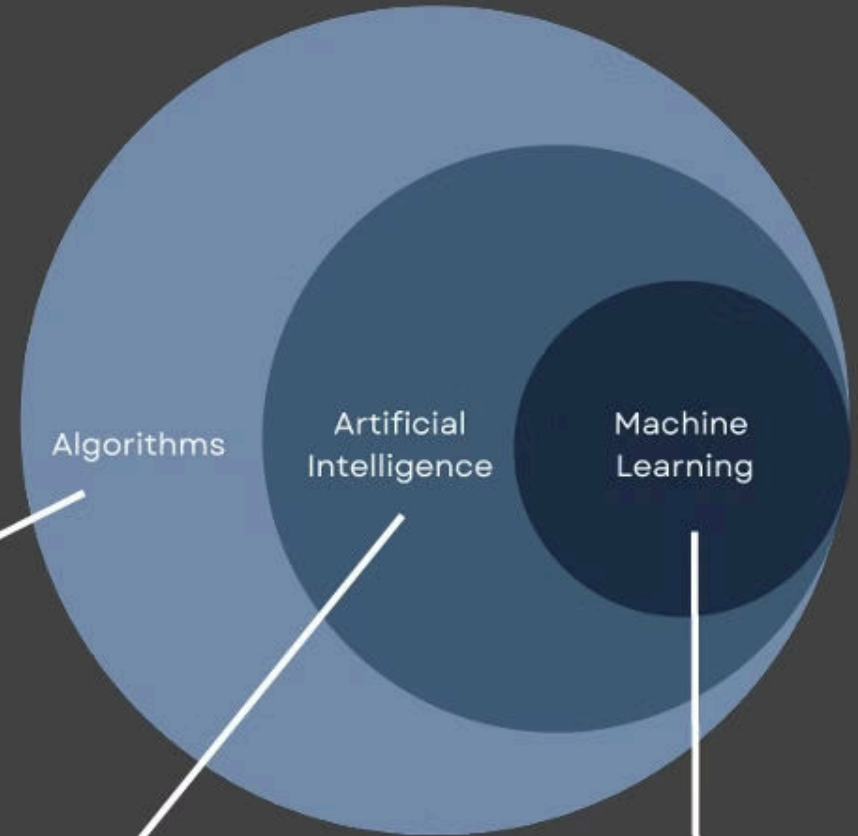


Source : WinWire via @BrianJohnson\_01

# Terminology in AI

AI for AI Ethics & Governance

# Nested Taxonomy



Process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

Algorithm(s) that able to perform tasks that normally require human intelligence

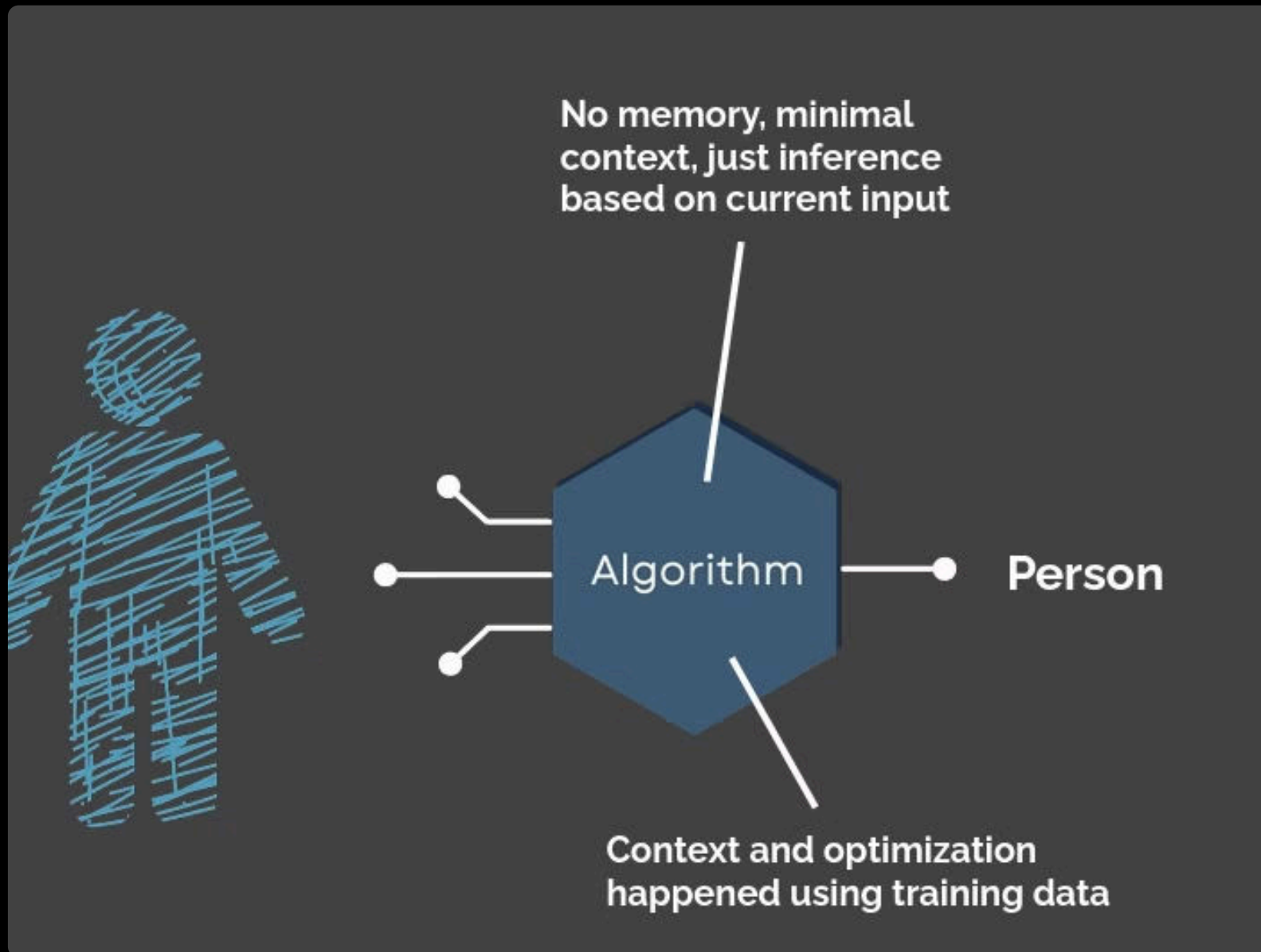
AI that adapts through experience



# Types of Artificial Intelligence Models - Reflexive Models

- **Nature:**
  - Focus on learning from interactions and adapting over time.
  - Often use feedback loops to refine their outputs based on new data.
- **Examples:**
  - Adaptive neural networks (including LLMs), reinforcement learning algorithms.
  - Dynamic environments where continuous learning is crucial, like financial market predictions.
- **Advantages:**
  - High adaptability to new data or environments.
  - Continuous improvement over time.
- **Challenges:**
  - Require ongoing data input for effective learning.
  - More complex to implement and govern due to their evolving nature.

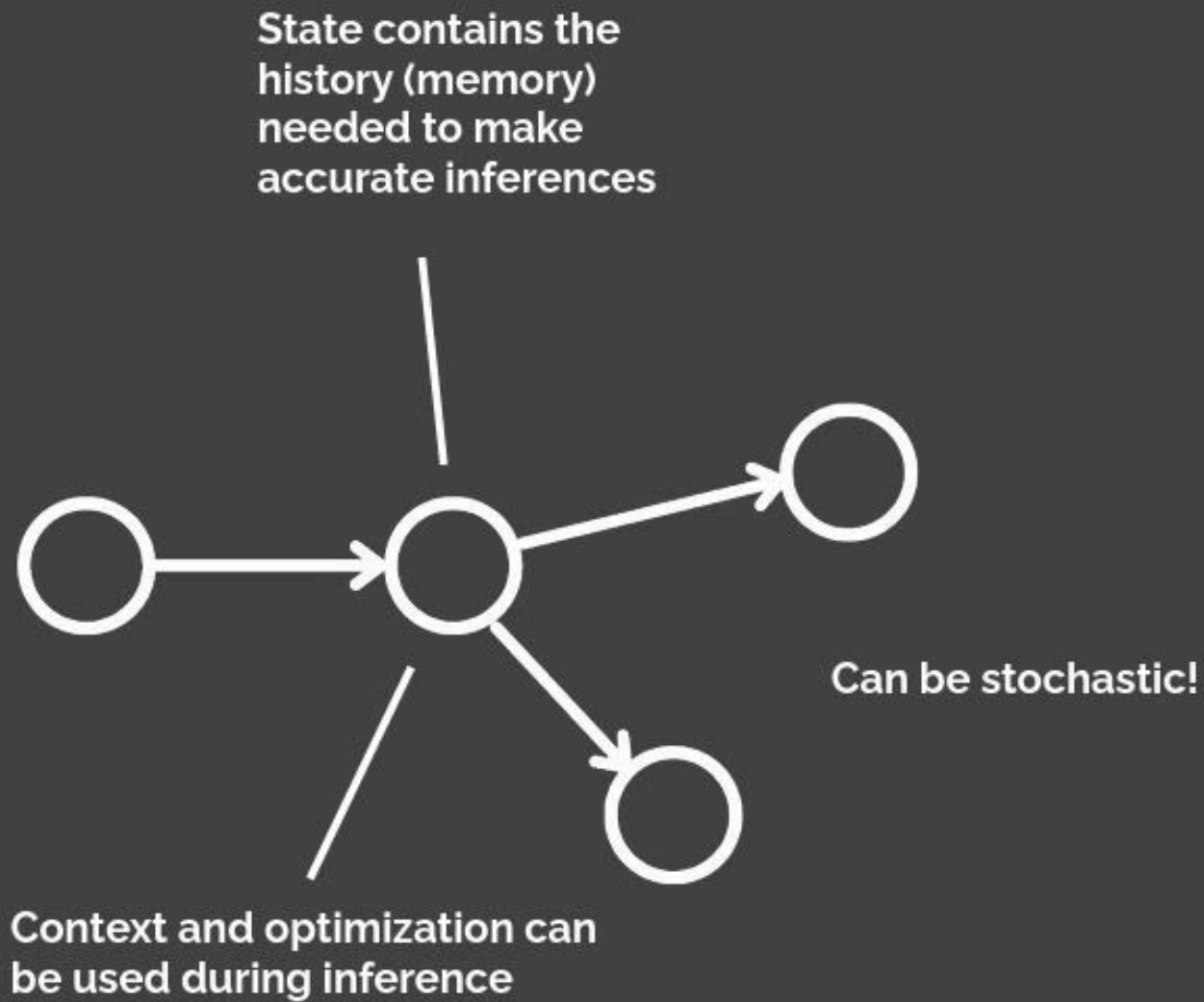
# Reflexive Models



# State-Based Models

- **Nature:**
  - Operate based on predefined states and transitions.
  - Use a set of rules and logic to determine outputs.
- **Examples:**
  - Finite-state machines, Markov decision processes.
- **Applications:**
  - Controlled environments with predictable states, like industrial automation.
- **Advantages:**
  - Predictable and explainable outputs.
  - Easier to govern and validate.
- **Challenges:**
  - Less flexible in adapting to unforeseen situations.
  - Limited by the predefined states and rules.

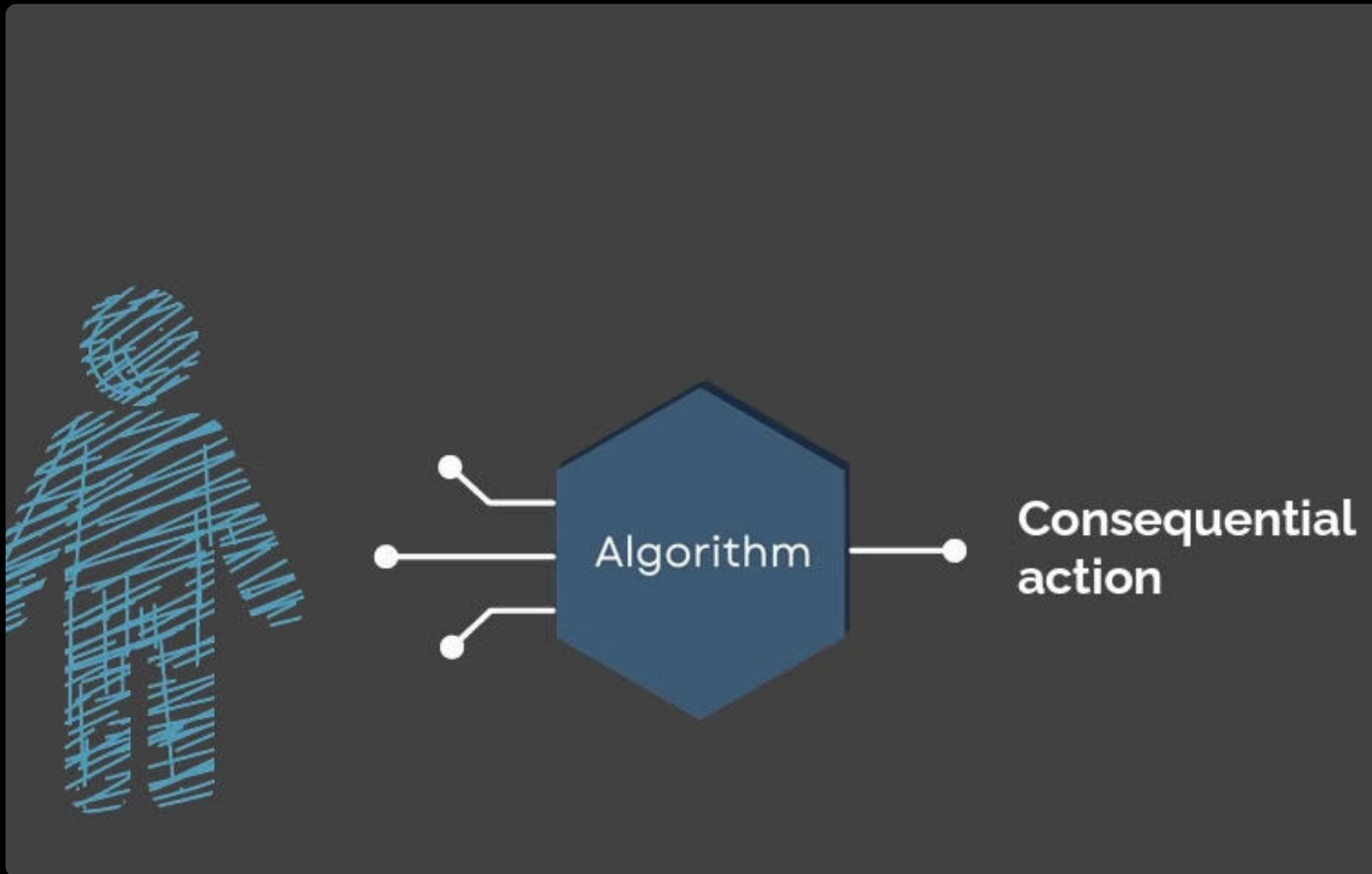
# State-Based Models



# Automated Decision Systems

- **Nature:**
  - Designed to make decisions or predictions without human intervention.
  - Can include a variety of AI models, including reflexive and state-based.
- **Examples:**
  - Credit scoring algorithms, automated hiring systems.
- **Applications:**
  - Decision-making in business, finance, HR, etc.
- **Advantages:**
  - Efficiency and speed in decision-making.
  - Can handle complex data sets and scenarios.
- **Challenges:**
  - Potential for bias and ethical concerns.
  - Requires careful governance to ensure fairness and transparency.

# Automated Decision Systems



# Governance Implications

- **Reflexive Models:**
  - Need for continuous monitoring and updating of models.
  - Challenges in ensuring consistent performance over time.
- **State-Based Models:**
  - Easier to set governance protocols due to predictability.
  - Must ensure the comprehensiveness of state definitions and transitions.
- **Automated Decision Systems:**
  - Necessity for thorough auditing and transparency to mitigate bias.
  - Importance of incorporating human oversight in critical decision points.

# Machine Learning - Types of Learning - Supervised Learning



## Regression



What will be the temperature tomorrow?

84°



Fahrenheit

## Classification



Will it be hot or cold tomorrow?

COLD

HOT

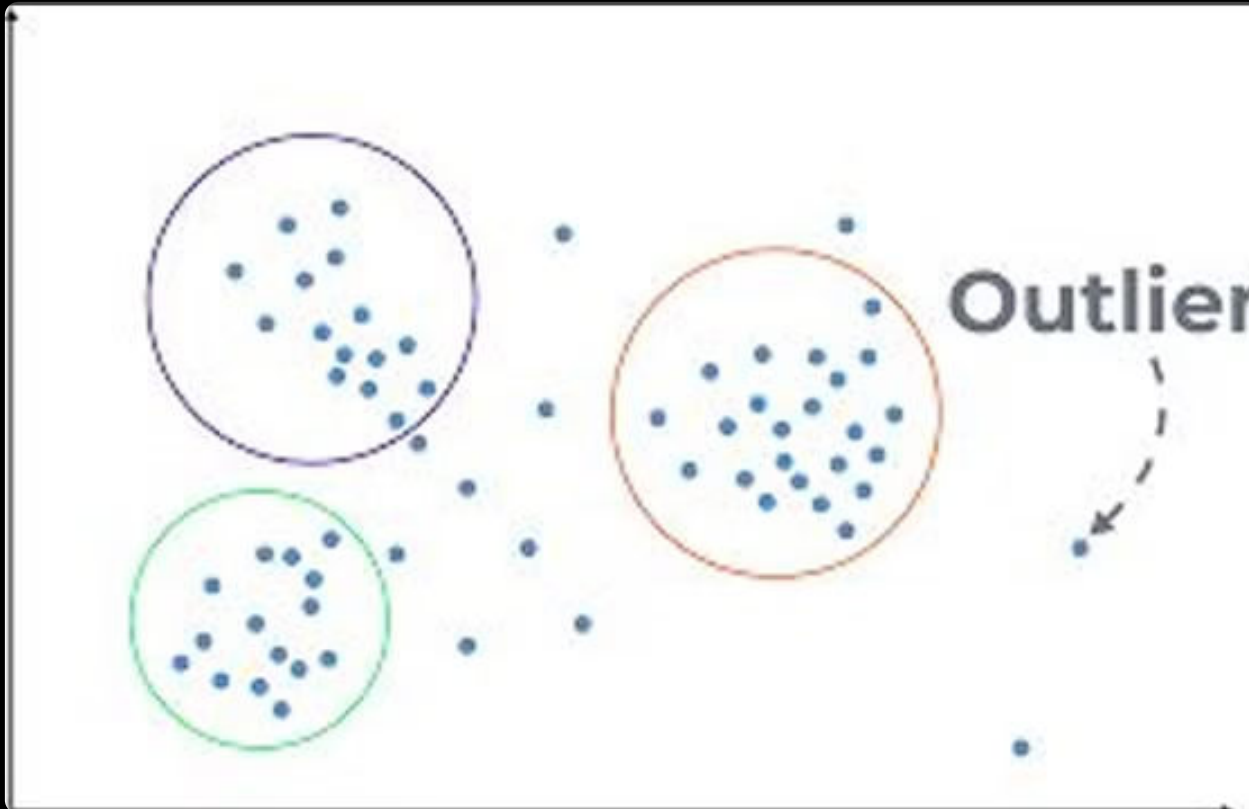


Fahrenheit



# Unsupervised Learning

- Clustering and Outlier Detection







# Reinforcement Learning

- No model of the world (to start), or an incomplete model of the world
- Learns through experience and rewards
- Goals
- Metrics of success

# Sources of Risk

## Data-Related Risks:

- **Data Quality and Bias:** Poor data quality or biased datasets can lead to inaccurate, unfair, or discriminatory outcomes. This includes biases in training data, which can propagate through the AI system.
- **Data Privacy and Security:** Lack of security measures, lack of Privacy Enhancing Technologies employed
- **Processing of Data itself**



## Algorithmic Risks:

- **Algorithmic Bias:** Algorithms can inherit or accentuate biases
- **Overfitting and Generalization:** AI models may overfit to training data, failing to generalize to new, unseen data.

# Sources of Risks (cont.)

## Technical Risks:

- **System Reliability and Robustness:** Risks of system failures, glitches, or unexpected behaviour, especially in critical applications.
- **Adversarial Attacks:** Vulnerability to adversarial attacks where small, often imperceptible changes to input data can lead to incorrect outputs.

## Operational Risks:

- **Integration and Scalability:** Challenges in integrating AI systems into existing processes and infrastructures.
- **Maintenance and Update:** Risks associated with maintaining and updating AI systems over time, including model drift.
- **Cybersecurity Vulnerabilities:** AI systems are susceptible to cybersecurity threats, with novel attacks (inference, extraction, system takeover)

# Sources of Risks (cont.)

## Human and Organizational Risks:

- **Human-AI Interaction:** Risks arising from the interaction between humans and AI systems, including over-reliance or misuse and lack of interaction.
- **Change Management:** Organizational risks related to adopting and adapting to AI technologies.

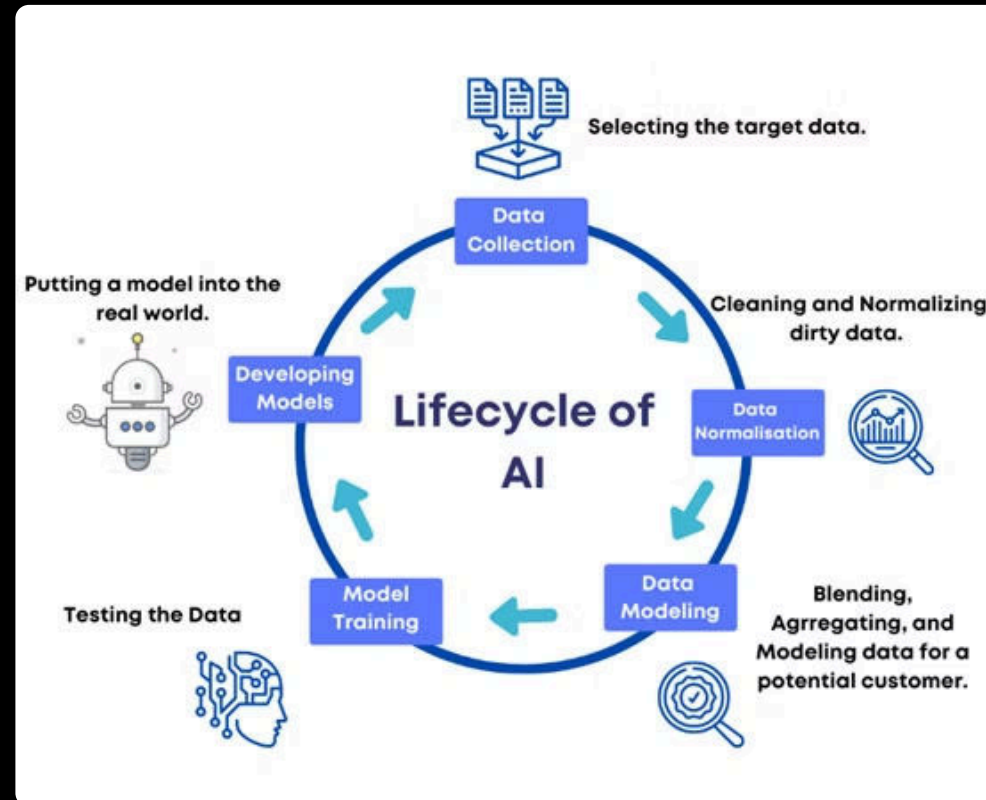
# Where is it riskier with AI? (Risk Drivers)

1. **For Bias: Scale and Speed of Impact/Opacity of Decision Making/Automated Decisions/Reinforcement of Existing Biases/Lack of Contextual Understandings**
2. **Rapid Technological Advancements:** The fast pace of AI development can outstrip the ability of organizations and regulatory bodies to keep up, leading to gaps in governance, standards, and ethical considerations.
3. **Interdependency and Integration Challenges:** AI systems are often integrated with other systems and technologies, making them highly interdependent. This can amplify risks, as a failure in one system can cascade to others.
4. **Regulatory and Compliance Changes:** The evolving legal and regulatory landscape around AI can be a risk driver, especially for cross-border operations where AI regulations may differ significantly.

- **Lack of Expertise and Understanding:** Insufficient knowledge or understanding of AI technologies among staff, management, or regulators can lead to mismanagement of AI risks.
- **Human Factors:** Over-reliance on AI decision-making, user errors, or resistance to adopting AI technologies can also drive risks.
- **Model Drift and Performance Issues:** Over time, AI models can drift from their original performance levels due to changes in the underlying data or environment, leading to decreased accuracy and reliability.
- **Scalability and Maintenance:** Challenges in scaling AI solutions and maintaining their performance over time can introduce risks, especially in dynamically changing environments.



# Risk can arise throughout the lifecycle of AI



# Example extract of an enumeration of cybersecurity risks during model development

**Table 7** Classification of threats to ML-based systems during development, where C, I, and A denote the loss of confidentiality, integrity, and availability, respectively.

| Threat                      | Sub-threat           | Damage   | Description   |
|-----------------------------|----------------------|--|---|
| T1.1 Data poisoning attack  | Malfunction          | I, A   | Manipulation of an ML data source or an ML dataset<br>- to cause a malfunction of the trained model<br>* for specific inputs<br>* for inputs that contain specific information<br>* for unspecified inputs                          |
|                             | Targeted             |  |   |
|                             | Backdoor             |  |   |
|                             | Non-targeted         | I, A   | - to obtain a trained model with an unintended functionality change   |
|                             | Functionality change |  |   |
|                             | Resource exhaustion  |  |   |
| Information embedding       | C                    | - to embed sensitive information into ML datasets to disclose it during system operation                           |   |
| T1.2 Model poisoning attack | Malfunction          | I, A   | Manipulation of a pre-trained model, a learning mechanism, or a trained model<br>- to cause a malfunction of a trained model<br>* for specific inputs<br>* for inputs that contain specific information<br>* for unspecified inputs |
|                             | Targeted             |  |   |
|                             | Backdoor             |  |   |
|                             | Non-targeted         | I, A   | - to obtain a trained model with an unintended functionality change   |
|                             | Functionality change |  |   |
|                             | Resource exhaustion  |  |   |
| Information embedding       | C                    | - to embed sensitive information into model parameters or hyperparameters to disclose them during system operation |   |

# ..and all related assets around the AI need to be assessed

**Table 3** List of assets in an ML-based system.

| Asset   | Description   |
|---|---|
| A6.1 Access control program                     | A program that controls the input of data for system operation.   |
| A6.2 Pre-processing program                     | A program that processes raw data to produce input to ML components. (This may access the ML components' internal information or may be combined with the ML components.)                                       |
| A6.3 ML component                               | A software component that implements a trained model and possibly its interpretation functionality.   |
| A6.4 Post-processing program                    | A program that processes the ML component's output and its interpretation.  |
| A6.5 Monitoring/risk treatment program          | A program that treats risks by monitoring the system's behavior.  |
| A6.6 Other conventional software components     | Other software components that do not consist of ML components.   |
| A6.7 System specification & related information | Information on the ML datasets, the trained models, the other system specifications, and their related information, such as datasets or models resembling the ones used in the system development or operation. |

**Table 4** List of assets in operation.

| Asset   | Description   |
|---|---|
| A7 Data source for system operation                                       | A population, a process, or an environment from which raw data instances are collected and used as input to ML-based systems. |
| A8 Data for system operation  | A set of data instances input to ML-based systems during operation.   |
| A9 Computing environment & operating organization during system operation | The computing environment used by the ML-based system and the organization that operates the system.                          |
| A10 System output   | Data that an ML-based system outputs.   |

Assets for the assessment also include data sources (training and operational) and conventional software components

# Risk Assessments (DPIAs) under the GDPR

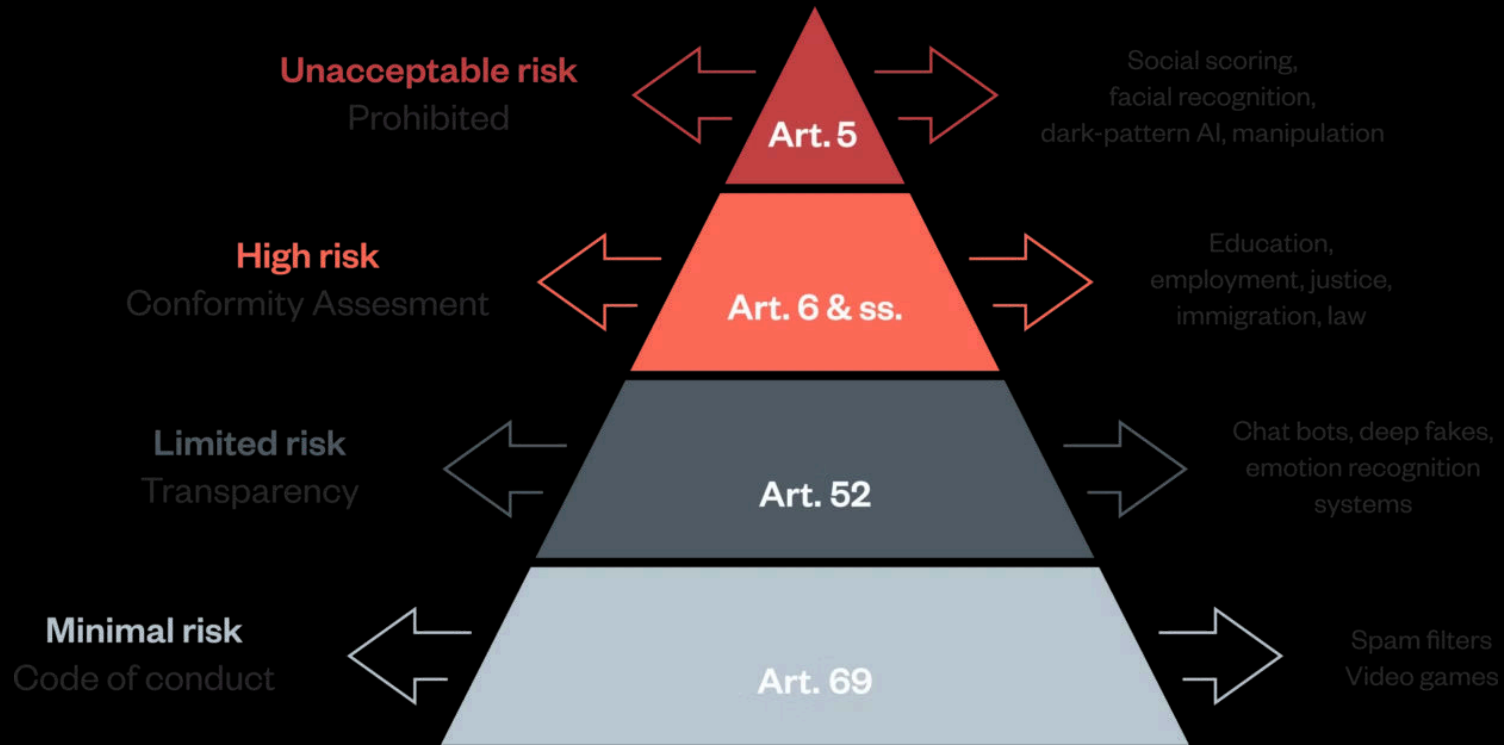
Data Protection Impact Assessments required when:

- **Systematic Profiling:** Extensive automated decision-making affecting individuals
- **Large-Scale Processing:** Processing personal data on a significant scale
- **Sensitive Data:** Handling special categories of sensitive personal data
- **Public Monitoring:** Systematic monitoring of public areas on a large scale
- **Cross-Border Transfers:** Sending data to non-EEA countries
- **Innovative Technology:** New tech with potential high privacy risks
- **Data Combination:** Merging data sources to create detailed profiles
- **Evaluation/Scoring:** Processing for evaluation or scoring purposes

# Automated Decision Making under GDPR

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.*

# EU Artificial Intelligence Act (AIA)



AI system is high-risk:

1. If it is used as a safety component of a product
2. OR if it is covered by one of 19 specified pieces of EU single market harmonization legislation (e.g., aviation, cars, medical devices)
3. OR AI systems deployed in the following sectors are deemed to be high-risk to safety or fundamental rights (e.g. critical infrastructure, employment, access to education...) → Domains (Education, Justice, Immigration) & purposes

# Conformity Assessments

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

Conformity assessment need to cover:

- data and data governance (including data protection)
- technical documentation
- record keeping
- transparency and provision of information to users
- human oversight
- robustness, accuracy, and security



# Why use DPIA for an AI assessment?

- AI Act complements the General Data Protection Regulation
- DPIAs are used to handling risks beyond privacy and are AI ready
  - Guidance for using DPIAs for AI systems already exists
- Introduction of a conformity assessment procedure under Article 43 for high-risk AI systems which requires assessment
  - DPIA can be used as the basis for a conformity assessment
- Need to assess data protection under Data Protection Impact Assessment (DPIA) in any event

There are some limitations of using the DPIA here, but it otherwise works well

# Limitations of DPIAs to Assess AI risks

## 1. **Complex AI Risks:**

- DPIAs may struggle to identify all risks in intricate AI systems
- AI's complexity makes risk identification challenging

## 2. **Technological Pace:**

- DPIAs might not keep up with rapid AI technological changes
- New risks can emerge faster than DPIAs can address

## 3. **Ethical Concerns:**

- DPIAs may not cover broader AI ethical issues
- Focus on data protection, so issues like fairness and bias might not be fully addressed

## 4. **Need remains to prioritize the risks**

Each assessment is different and drives the choice of tools

Other artefacts are needed to combine with and complement the DPIA

# Guidance on AI and Data Protection

ICO guidance covers topics such as transparency, accountability, and lawful basis for processing. It also covers how to implement data protection by design and default, and how to carry out data protection impact assessments (DPIAs) (updated for AI).

For more information, visit the ICO's website: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>

The logo for the Information Commissioner's Office (ICO) features the lowercase letters 'i', 'c', and 'o' in a bold, dark blue font. The 'i' has a solid blue dot above it. The 'c' and 'o' are also in a bold, dark blue font. A solid blue dot is positioned to the right of the 'o', serving as a period. The entire logo is centered on a white background with a light gray checkerboard pattern.

Information Commissioner's Office

# Complementary Tools

See Catalogue of Tools & Metrics for Trustworthy AI from OECD.AI



# Risk Assessment Template

The image shows a Google Sheets spreadsheet titled "Risk-Impact Assessment". The spreadsheet is organized into columns A through H. Column A is labeled "Stakeholder", B is "Interest/Harm", C is "Metric/Cause", D is "Likelihood", E is "Likelihood Notes", F is "Magnitude", G is "Magnitude Notes", and H is "Risk Level". The rows are numbered 1 through 20. The data is as follows:

| 1  | A           | B             | C             | D          | E                | F         | G               | H          |
|----|-------------|---------------|---------------|------------|------------------|-----------|-----------------|------------|
| 1  | Stakeholder | Interest/Harm | Metric/Cause  | Likelihood | Likelihood Notes | Magnitude | Magnitude Notes | Risk Level |
| 2  | Candidate   |               | Transparency  | Low        |                  | Medium    |                 | Medium     |
| 3  |             |               | Effectiveness | Medium     |                  | High      |                 | High       |
| 4  |             |               | Misuse        | High       |                  | Low       |                 | Low        |
| 5  |             |               |               |            |                  |           |                 |            |
| 6  |             |               |               |            |                  |           |                 |            |
| 7  |             |               |               |            |                  |           |                 |            |
| 8  | Developer   |               |               |            |                  |           |                 |            |
| 9  |             |               |               |            |                  |           |                 |            |
| 10 |             |               |               |            |                  |           |                 |            |
| 11 |             |               |               |            |                  |           |                 |            |
| 12 | Deployer    |               |               |            |                  |           |                 |            |
| 13 |             |               |               |            |                  |           |                 |            |
| 14 |             |               |               |            |                  |           |                 |            |
| 15 |             |               |               |            |                  |           |                 |            |
| 16 | Society     |               |               |            |                  |           |                 |            |
| 17 |             |               |               |            |                  |           |                 |            |
| 18 |             |               |               |            |                  |           |                 |            |
| 19 |             |               |               |            |                  |           |                 |            |
| 20 |             |               |               |            |                  |           |                 |            |

Source: Babl.ai

# AI Impact Assessment Template Aligned with ISO42005

## Section A: System Information

- AI System Description
- Purpose and Objectives

## Section B: Stakeholder Identification

- Relevant Interested Parties

## Section C: Impact Analysis

- Potential Benefits
- Potential Risks and Harms

## Section D: Data Governance

- Data Types and Sources
- Data Quality and Integrity

## Section E: Algorithmic Accountability

- Algorithm Description
- Impact of Algorithms

## Section F: Compliance and Monitoring

- Regulatory Adherence
- Ongoing Monitoring and Evaluation

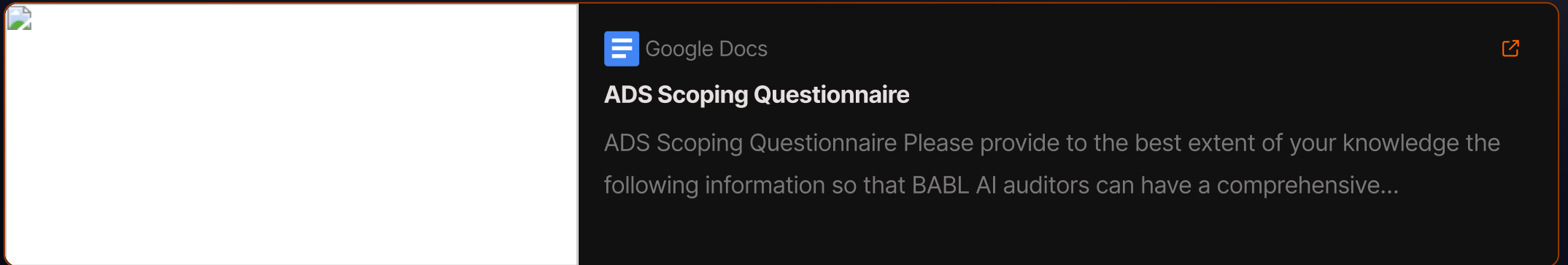
## Section G: Documentation and Reporting



- Assessment Documentation
- Disclosure and Transparency

## Final Notes

- Approval and Revision:
- Ethical Considerations

# Scoping Questionnaire



 Google Docs 

**ADS Scoping Questionnaire**


ADS Scoping Questionnaire Please provide to the best extent of your knowledge the following information so that BABL AI auditors can have a comprehensive...

Source: Babl



# Contextual Tools



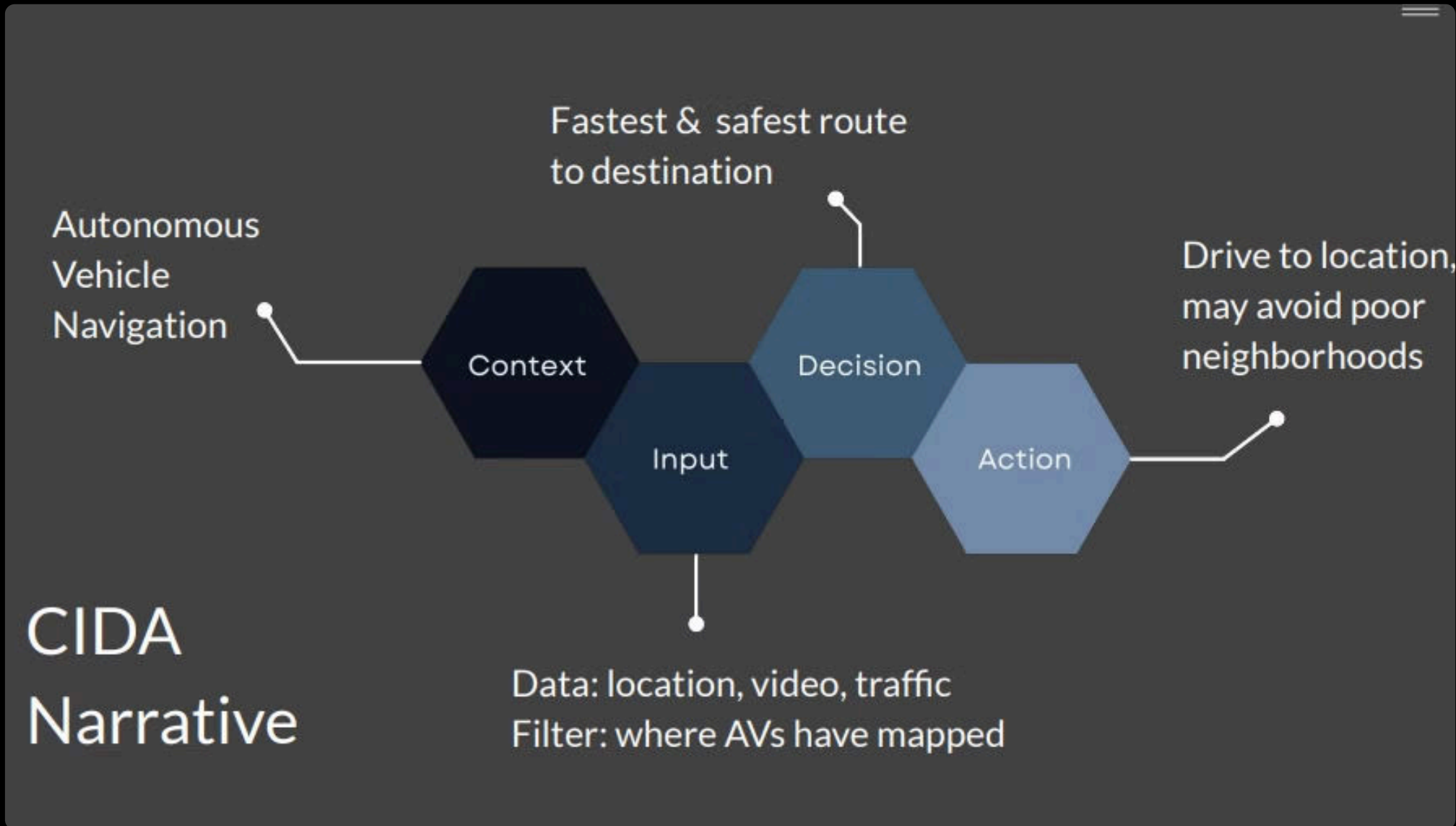
 BABL AI on Notion



## **Risk Assessment Requirements - Question List**

These questions form the basic contextual requirements to undergo a formal Algorithmic Risk and Impact Assessment.

# Context, Input, Decision and Action (CIDA)



Source: Babl

# Catalogues of Risks

## General

PLOT4ai



### **PLOT4ai - Assessments - Quick Check**

A threat modeling library to help you build responsible AI



Source: Plot4.ai

## Cybersecurity

**MITRE | ATLAS™**

MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems), is a knowledge base of adversary tactics, techniques, and case studies for machine learning (ML) systems based on real-world observations, demonstrations from ML red teams and security groups, and the state of the possible from academic research.

Source: the Mitre Corp.

# Ethics Tools

## Value Statements

### Our AI project:


**1** Aims to support the right / freedom / ability for persons [\*] to do [\*]

**2** Can trigger challenges with respect to the right / freedom / ability for persons [\*] to do [\*]

**3** Generally, the expected and desired outcome is that[\*]

# Human Rights Impact Assessment Guidance and Toolbox



 The Danish Institute for Human Rights




## Human rights impact assessment guidance and toolbox

Guidance and practical tools for conducting, commissioning, reviewing and monitoring human rights impact assessments of business projects.

# Bias Tooling

Data Science and Public Policy  
Carnegie Mellon University

 Data Science and Public Policy

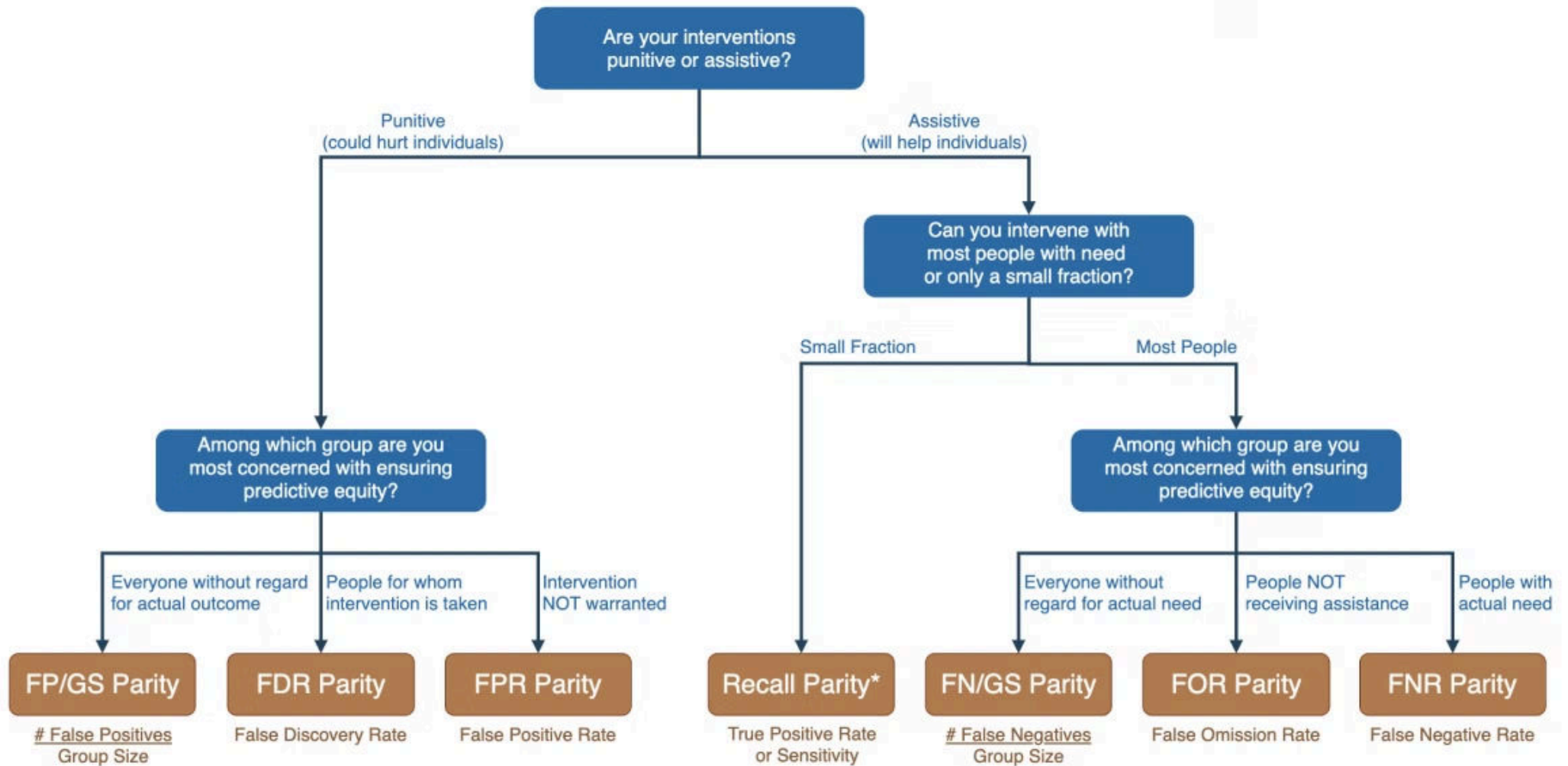


## Aequitas

Aequitas An open source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discriminati...





# Fairness Tree (Bias)

## FAIRNESS TREE (Zoomed in)



# The Top Artefacts for Assessments!

The top-notch artefacts:

-  Context Input Decision Action (CIDA)
-  Data Protection Impact Assessment (AI)
-  Risk Assessment Template
-  Value Statements



allc