SOBIGDATA.it

ITALIAN RESEARCH INFRASTRUCTURE

Bias Whisperers: The Art of Detecting Subtle Prejudices in AI Systems

Angelica Marotta

MIT Sloan School of Management

Pisa, May 13th, 2025

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE

Consiglio Nazionale delle Ricerche

About me

- **Profile:**
- **Researcher at MIT Sloan School of Management**
- **PhD in Cybersecurity**
- **Academic Expert | Speaker | Consultant**
- **Passionate about bridging the gap between research & industry**

- **Research and Work Areas:**
- **Harmonization of cybersecurity policies & global regulations**
- **AI security, compliance & ethical risks**
- **Cybersecurity culture & organizational resilience**
- **Cyber risk management & cyber insurance**

Research Focus:

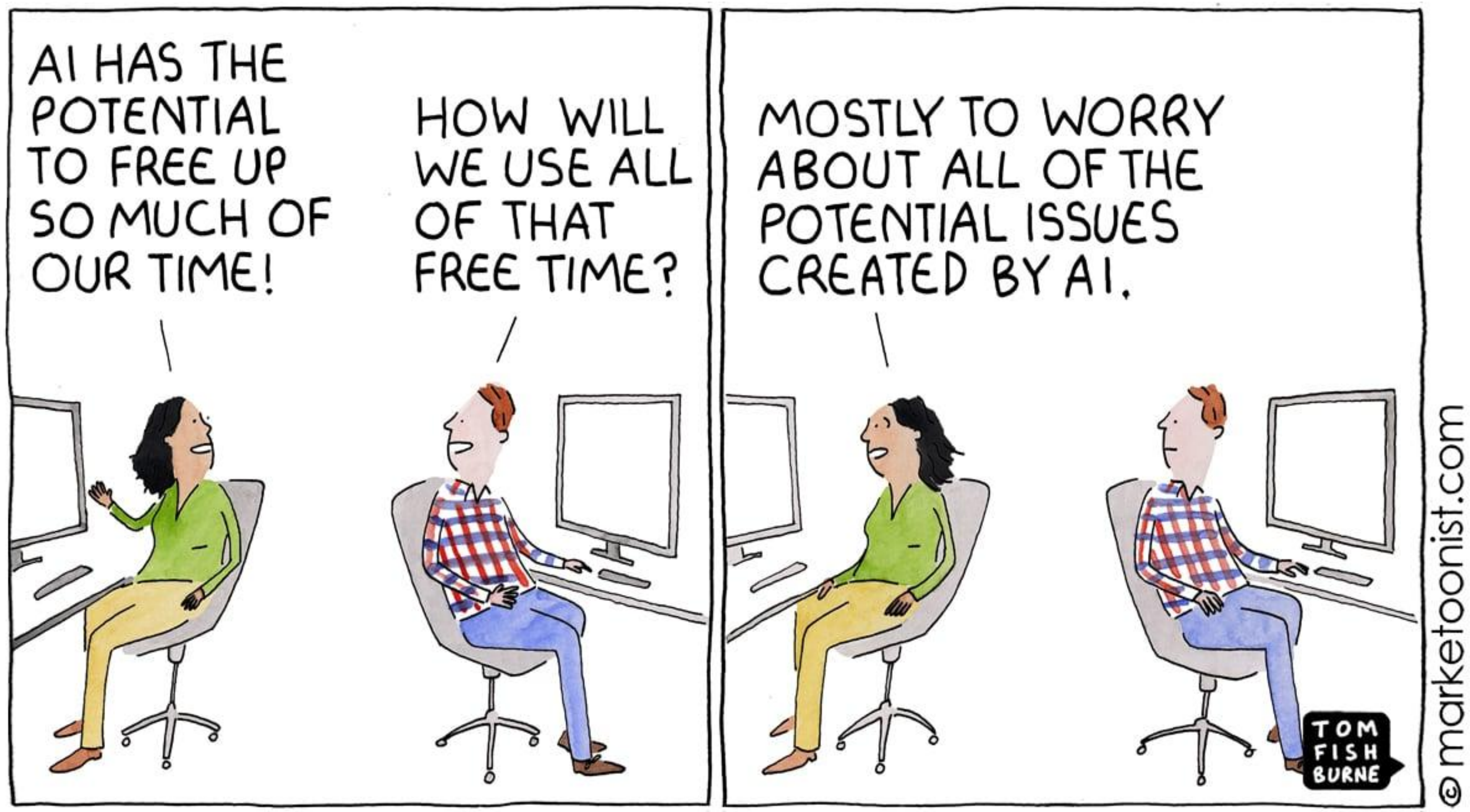
- **Ethical AI implementation (especially in healthcare), ensuring data security and regulatory compliance**

SOBIGDATA.it ITALIAN RESEARCH INFRASTRUCTURE

Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani Piano Nazionale di Ripresa e Resilienza | SOBIGDATA.it Italian Research Infrastructure | Consiglio Nazionale delle Ricerche

The AI Paradox: Time Saved vs. New Problems Created

Q: “In your research, have you encountered a technological solution that solved one problem but unexpectedly created another? How did you address this paradox?”


AI HAS THE POTENTIAL TO FREE UP SO MUCH OF OUR TIME!
HOW WILL WE USE ALL OF THAT FREE TIME?
MOSTLY TO WORRY ABOUT ALL OF THE POTENTIAL ISSUES CREATED BY AI.
TOM FISH BURNE
© marketoonist.com

AI Bias: The Elephant in the Room

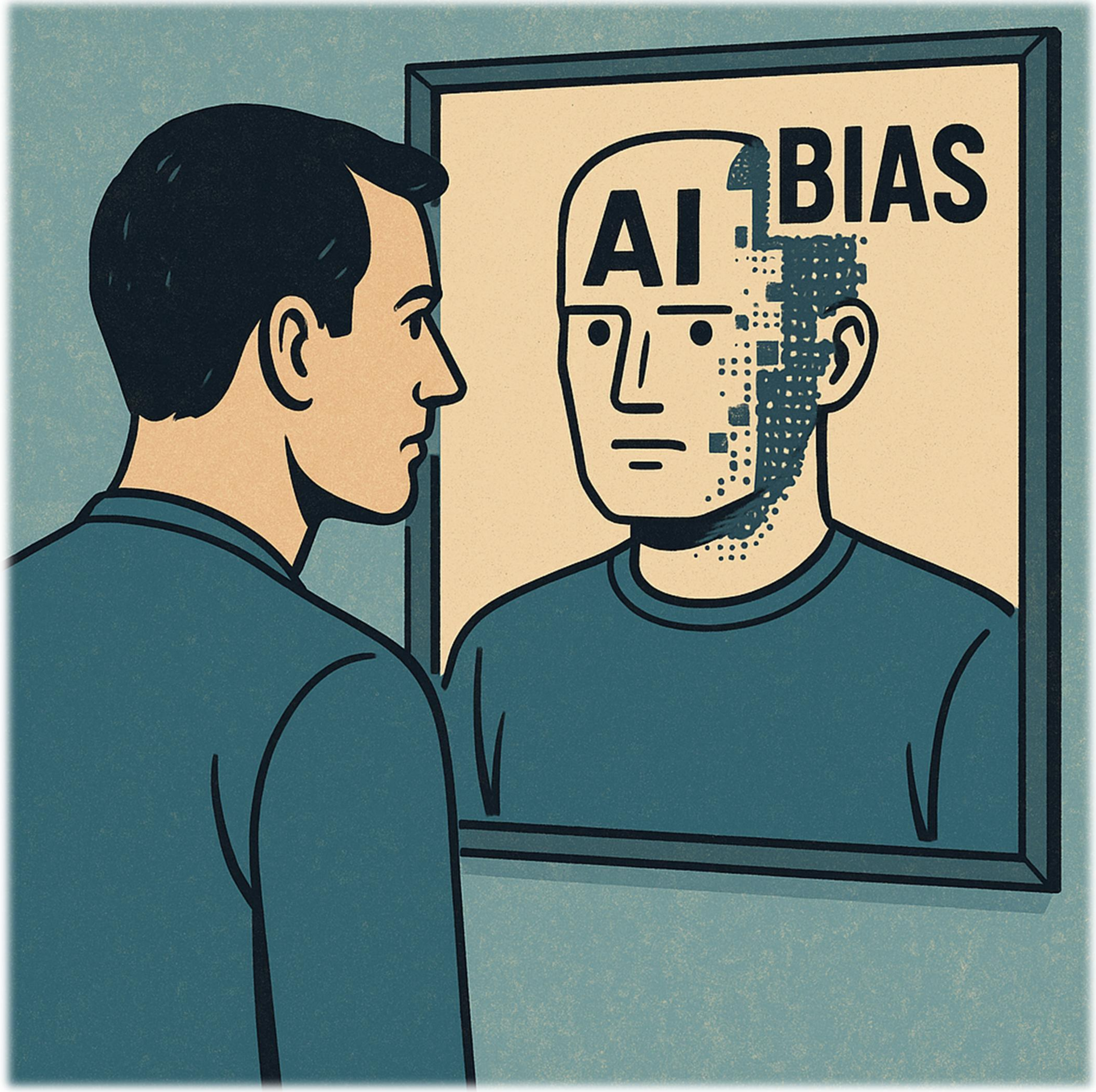
We create AI to solve problems and save time…but end up creating new challenges that demand our attention…

AI bias is perhaps the most critical of these challenges:

- **Systems that mirror and amplify human prejudices**
- **Algorithms that favor certain groups over others**
- **Technologies that can entrench existing inequalities**

AI Bias…is It a Tech Issue or A Human Challenge?

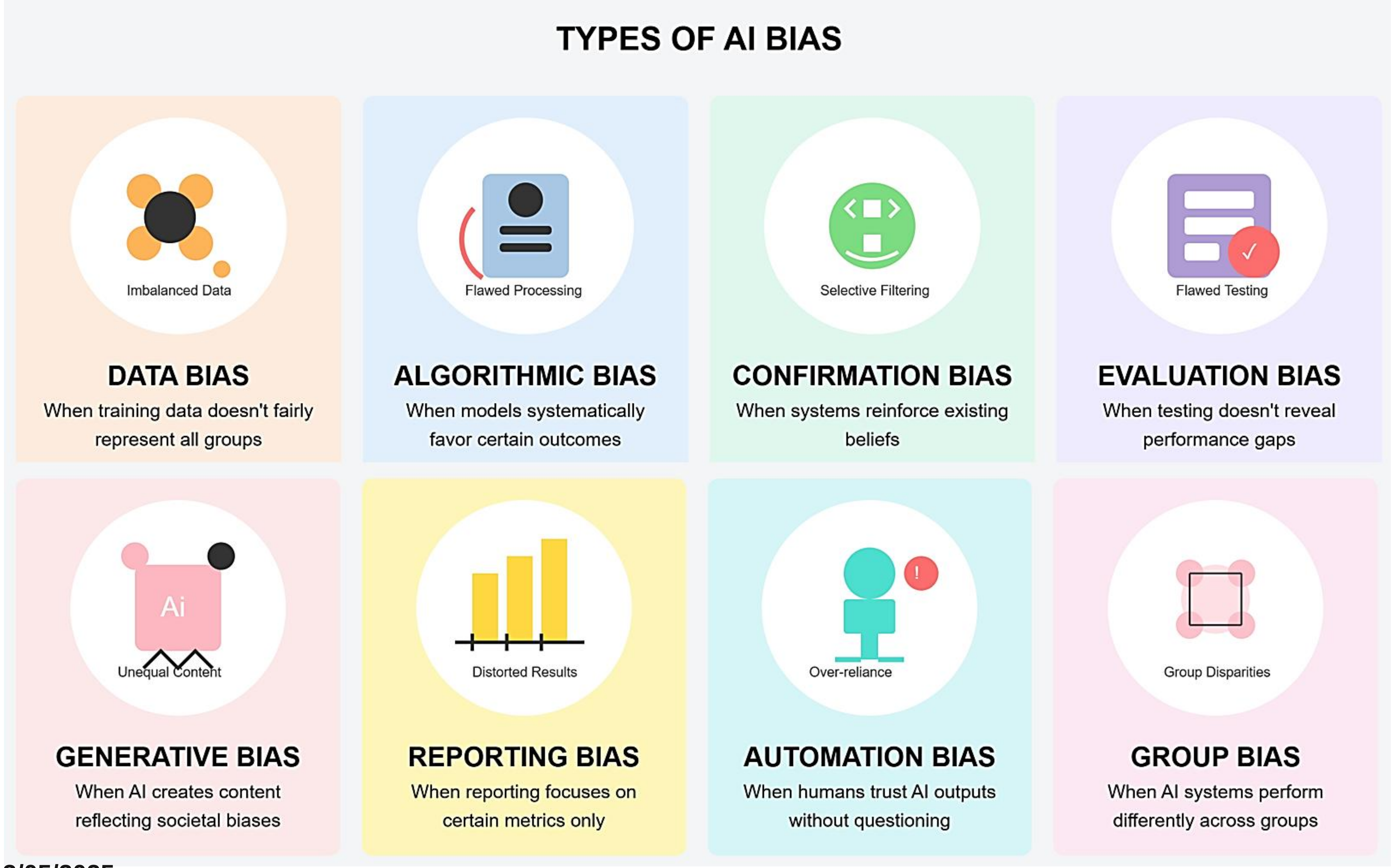
Imagine AI as a mirror, not just refelecting technology capabilities, but also reflecting our existing social structures



© A. Marotta, 2025 – 13/05/2025

Understanding AI biases

AI bias occurs when an automated system consistently produces unfair or prejudiced results due to flawed algorithms, skewed training data, or problematic design choices.



TYPES OF AI BIAS

DATA BIAS
When training data doesn't fairly represent all groups

ALGORITHMIC BIAS
When models systematically favor certain outcomes

CONFIRMATION BIAS
When systems reinforce existing beliefs

EVALUATION BIAS
When testing doesn't reveal performance gaps

GENERATIVE BIAS
When AI creates content reflecting societal biases

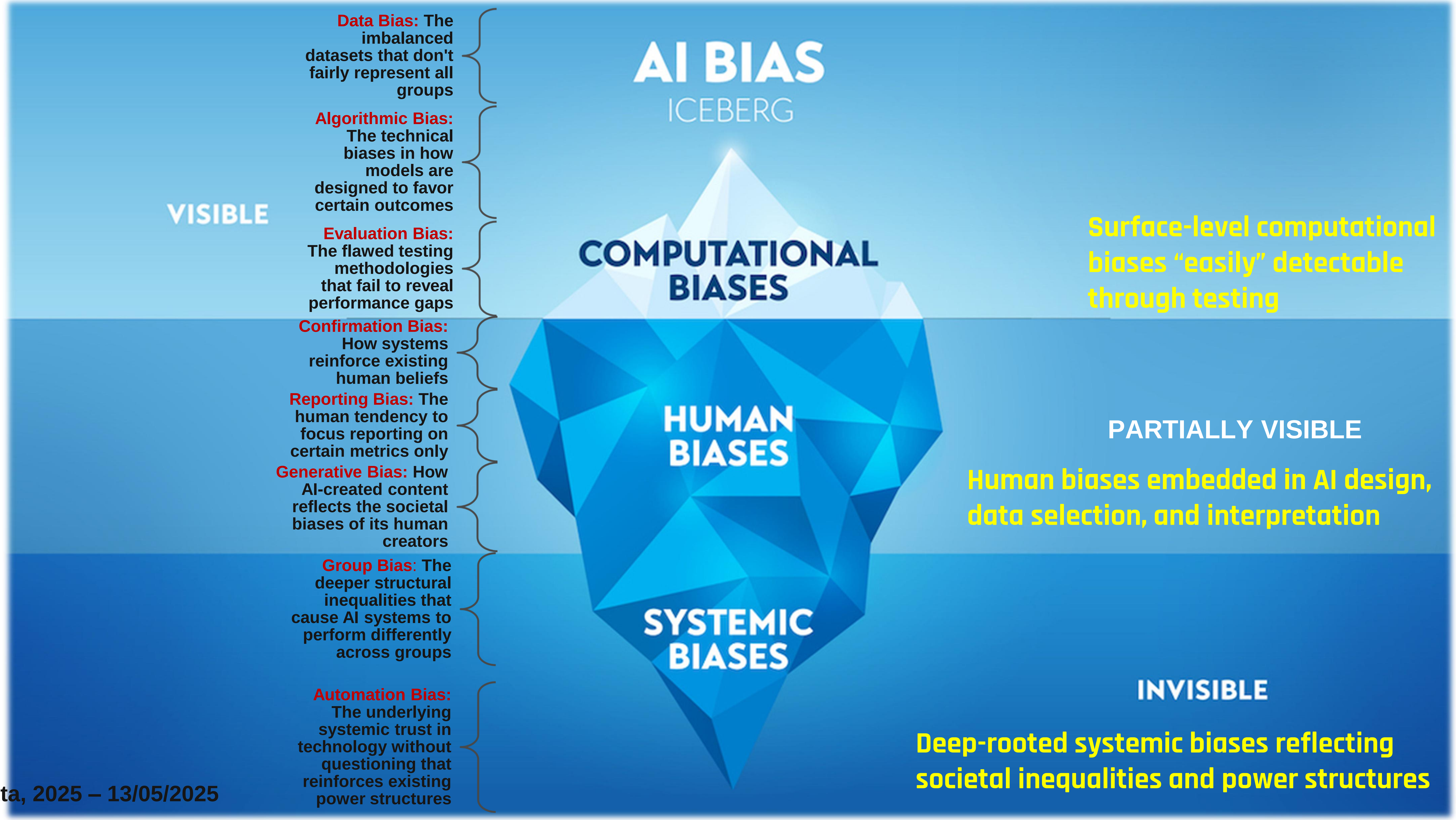
REPORTING BIAS
When reporting focuses on certain metrics only

AUTOMATION BIAS
When humans trust AI outputs without questioning

GROUP BIAS
When AI systems perform differently across groups

The Hidden Depths of AI Bias?

AI bias manifests at multiple levels, with visible computational issues merely the tip of an iceberg concealing deeper human and systemic biases that shape technology's impact on society.



Data Bias: The imbalanced datasets that don't fairly represent all groups

Algorithmic Bias: The technical biases in how models are designed to favor certain outcomes

Evaluation Bias: The flawed testing methodologies that fail to reveal performance gaps

Surface-level computational biases “easily” detectable through testing

Confirmation Bias: How systems reinforce existing human beliefs

Reporting Bias: The human tendency to focus reporting on certain metrics only

Generative Bias: How AI-created content reflects the societal biases of its human creators

PARTIALLY VISIBLE

Human biases embedded in AI design, data selection, and interpretation

Group Bias: The deeper structural inequalities that cause AI systems to perform differently across groups

Automation Bias: The underlying systemic trust in technology without questioning that reinforces existing power structures

INVISIBLE

Deep-rooted systemic biases reflecting societal inequalities and power structures

AI Bias in Healthcare

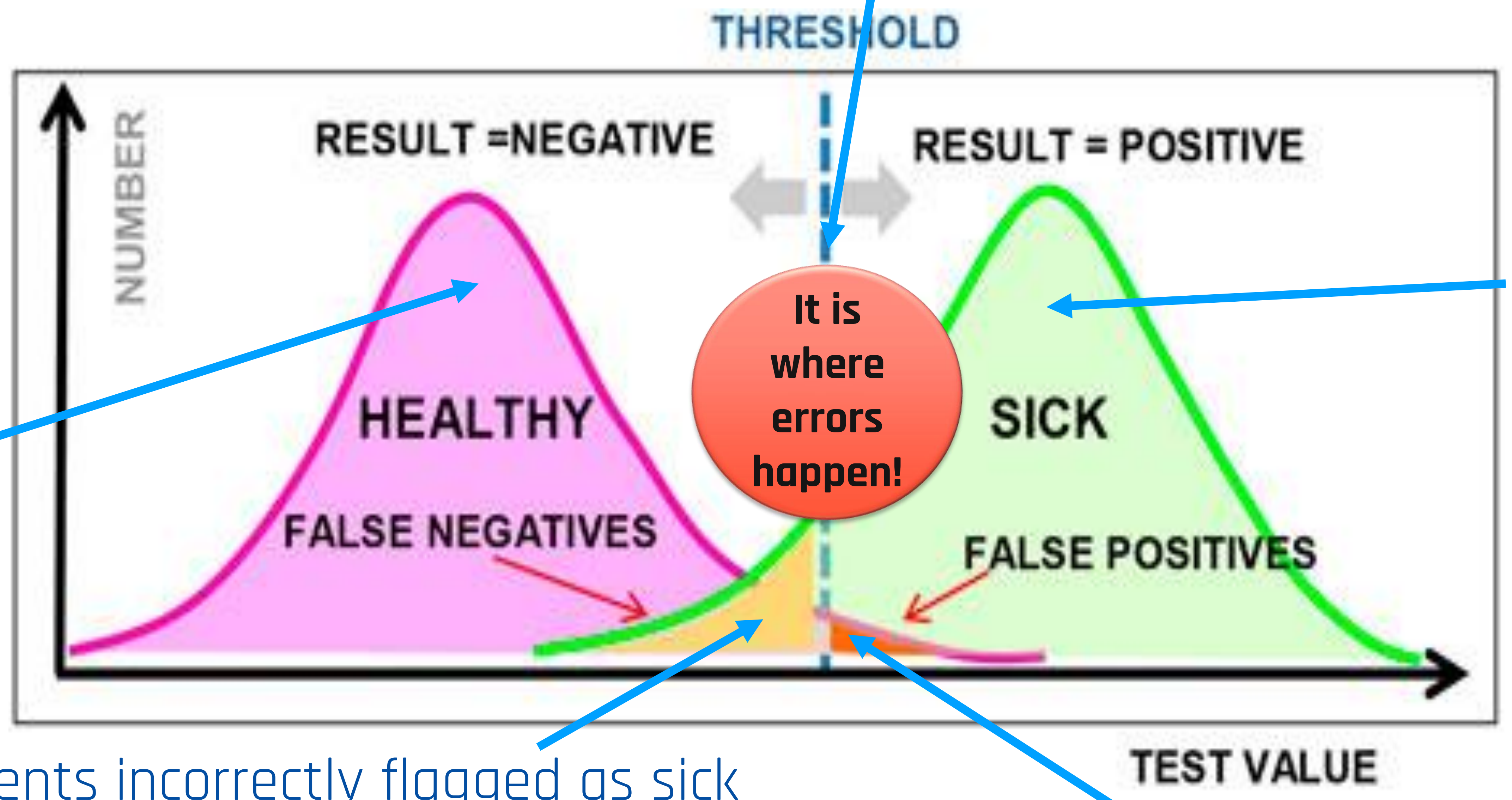
WHY IT MATTERS

Healthcare AI presents one of the highest-stakes domains for bias, where **algorithmic unfairness** can directly impact **patient outcomes**, **quality of care**, and even **survival rates**.

- **Life-critical decisions** are increasingly guided by AI systems
- **Historical biases** in medical research and practice become encoded in algorithms
- **Vulnerable populations** face compounded disadvantages when bias affects care
- **Regulatory oversight** struggles to keep pace with rapid AI healthcare deployment

AI uses this line to make yes/no decisions

AI diagnostic thresholds directly impact patient outcomes through **classification errors**



It is where errors happen!

Diagnosed as healthy

Diagnosed as sick

Healthy patients incorrectly flagged as sick (missing actual disease)

- **Example:** AI systems misses 20% more heart attacks in women than men because they were trained to recognize "typical" male symptoms like chest pain

Healthy patients incorrectly flagging healthy patients

- **Example:** AI misinterprets luteal phase estradiol decrease and mild mood shift as depression or cognitive issues, disregarding normal menstrual cycle fluctuations.

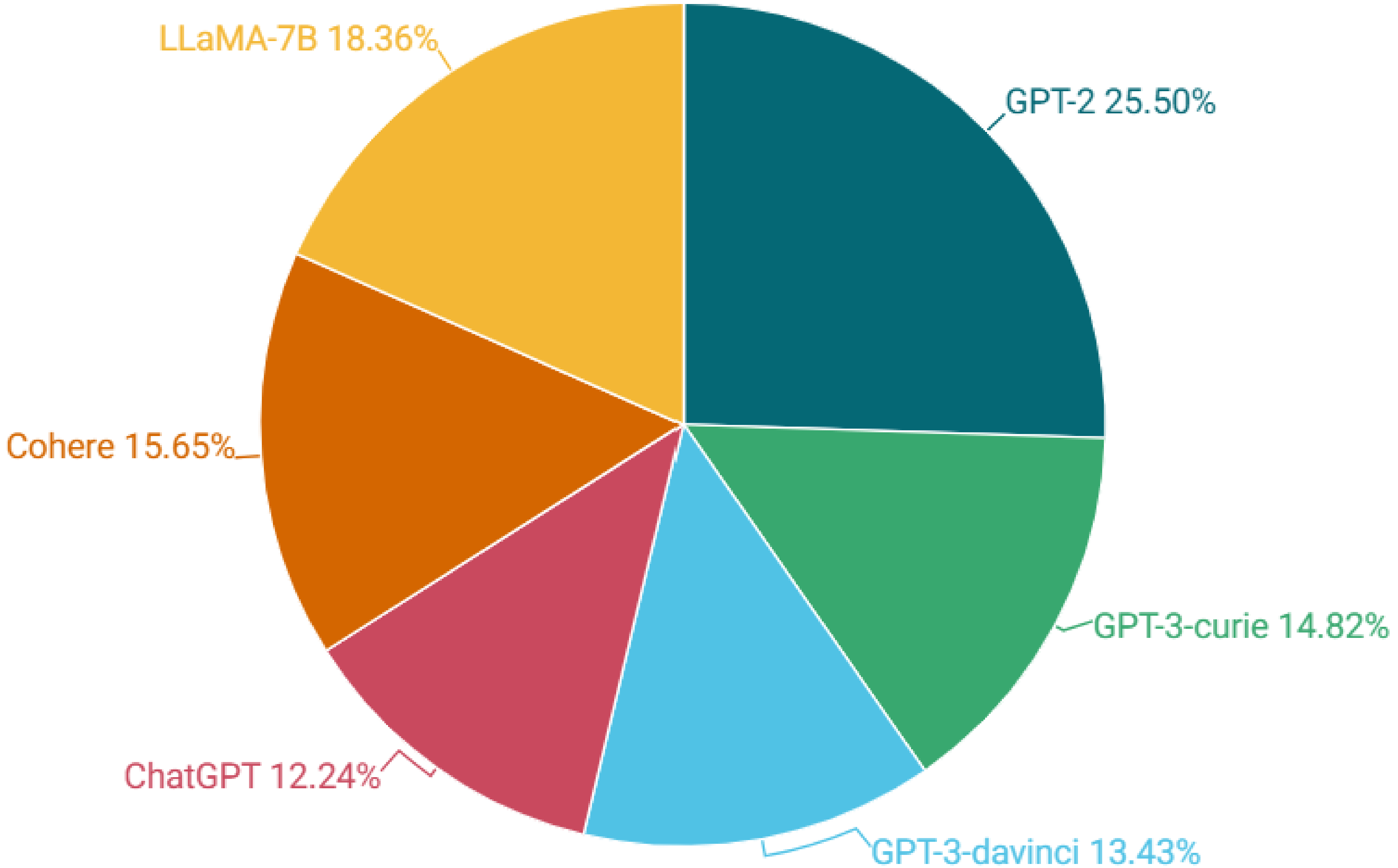
AI biased data → optimization for populations most represented in the training data

AI Bias in *Women*'s Healthcare

Women have historically been underserved by medical research and practice, and AI systems are now amplifying these **disparities** in alarming ways.

- **Women's health issues** dismissed for centuries as "hysteria" or psychosomatic
- **Clinical trials** predominantly male until policy changes in 1993, creating decades of biased data
- **Female-specific conditions** understudied, leading to diagnostic gaps AI cannot overcome
- **Medical manifestations** reported by women taken less seriously by both human and AI diagnosticians
- Women may receive less accurate **diagnoses**, inappropriate **medications**, and lower **quality care** from AI systems trained on male-centric medical knowledge.

Gender bias scores across six leading large language models (LLMs):



Source: 2024 Nature study

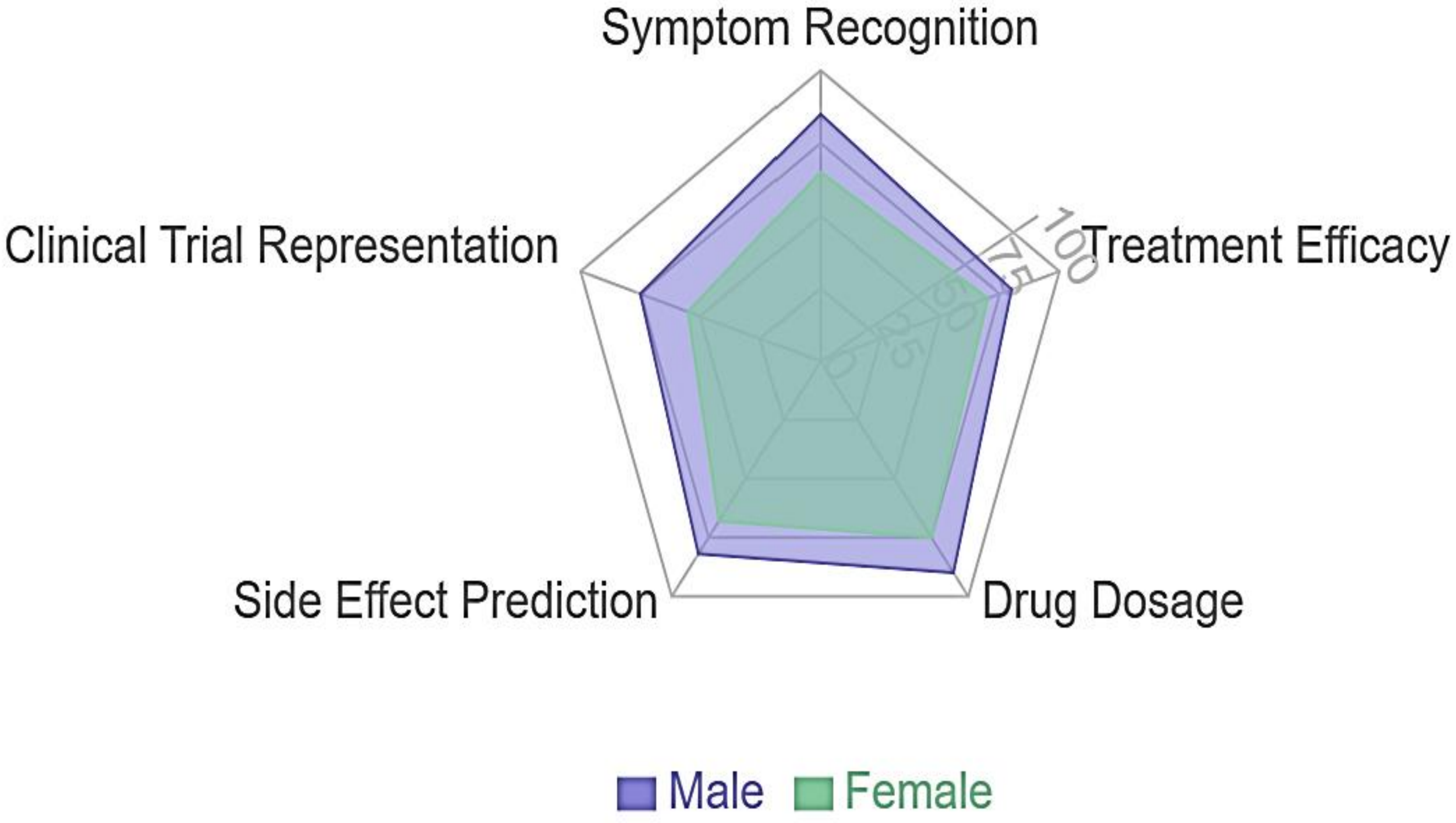
- All six leading LLMs exhibit **significant gender bias**, with **GPT-2** showing the highest at 25.50%, while even the "best" performer (**GPT-3-davinci**) still displays substantial bias at 13.43%.
- The **persistent bias** across diverse model architectures from different companies (OpenAI, Meta, Anthropic) suggests this is a **systemic industry-wide problem** rather than an isolated issue.

Gender Bias in AI Healthcare

Each type of bias manifests in real-world applications, revealing how biases at all levels of the "iceberg" converge to create measurable disparities in care.

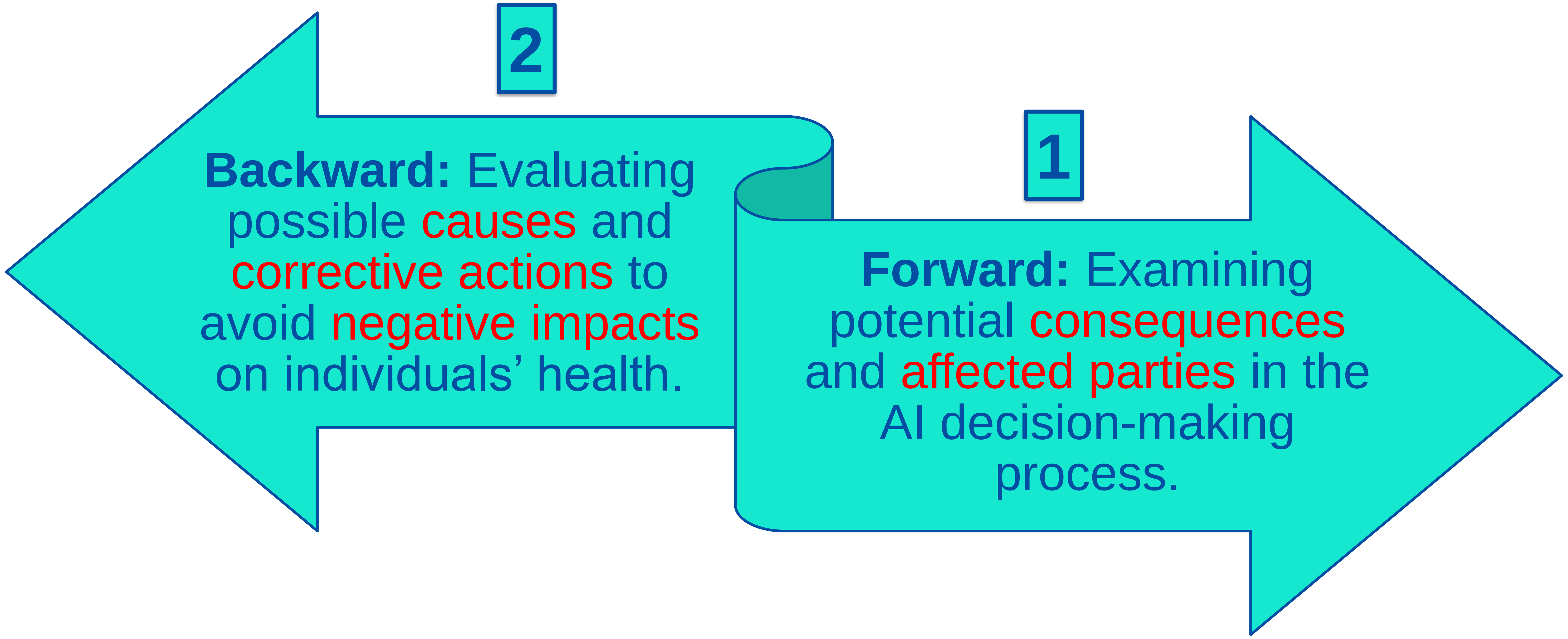
Example: First iceberg layer → Computational Biases (Visible):

- Measurable disparities in symptom recognition **(Evaluation Bias)**
- Quantifiable differences in drug dosage recommendations and female-specific side effects predictions **(Algorithmic Bias)**
- Documented bias in treatment efficacy predictions **(Data Bias)**



A method for analyzing AI bias issues

Forward-Backward Approach: Splitting the analysis of a problem into two opposite directions, “forward” and “backward,” in such a way as to cover multiple perspectives (attacking a problem from both sides).


2

Backward: Evaluating possible causes and corrective actions to avoid negative impacts on individuals’ health.

1

Forward: Examining potential consequences and affected parties in the AI decision-making process.

This method provides a **multidimensional view** of AI decision-making, helping to identify emerging obstacles and divergences between theory and practice.

Forward Perspective: Who Should Be Liable for Mistakes

The AI decision, influenced by inherent biases, raises **liability questions** for potential mistakes

Cause of Damage	Type of Liability	Stakeholder (Who is Liable)
Damage caused when AI was in use	Medical malpractice, vicarious liability	First category (Clinicians, medical staff)
Damage traced back to the design, implementation, or production of the AI system	Product liability	Second category (AI developers, manufacturers)

Medical Malpractice

Clinicians failing to assess AI recommendations critically ("a failure to exercise due care").

Example: Misdiagnosis due to incorrect AI output

Vicarious Liability

Institutions responsible for employee errors with AI.

Example: Nurse misinterpreting AI results

Product Liability

AI developers/manufacturers responsible for system errors.

Example: Biased training data leading to incorrect decisions

Backward Perspective: Behind the Algorithm

- AI systems often act as “black boxes” with hidden reasoning

- Bias can infiltrate at every stage of the AI data lifecycle in healthcare. Addressing these biases requires vigilance throughout the entire process

Historical Data Aggregation

- Source of bias (**Systemic Bias**): Societal prejudices
- **Example:** Studies on HIV and Autism predominantly contain data on male patients

Data Collection

- Source of bias (**Selection Bias**): Personal biases of data collectors
- **Example:** Unconscious selectivity against women in treatment data gathering

Data Processing

- Source of bias (**Algorithmic Bias**): Lack of context understanding
- **Example:** Misinterpretation of weight measurements without considering gender differences

Data Analysis & Management

- Source of bias (**Evaluation Bias**): Statistical bias
- **Example:** Analysis of electronic health records (EHRs) lacking information on women

Data Visualization

- Source of bias (**Reporting Bias**): Unconscious biases in data labeling
- **Example:** Mislabeling critical indicators due to limited exposure to female-specific symptoms

Data Storage

- Source of bias (**Confirmation Bias**): Insecure storage of biased data
- **Example:** Unencrypted storage leading to potential exploitation by cyber attackers

Data Utilization

- Source of Bias (**Generative Bias**): Biased decision-making in healthcare applications
- **Example: AI** misdiagnosing heart attacks in women due to bias

Q: Which phase is the most problematic?

A systematic process for evaluating and monitoring AI systems

Implementation of auditing processes (Algorithm Auditing Framework)

Focus on three key areas (with examples):

Liability: Ensuring compliance with relevant laws and regulations.	if a company operates in the EU needs to use algorithms that do not consider factors, such as gender, race, or religion (GDPR).
Technique: Employing necessary mechanisms for data security, protection, and explainability.	Algorithm designers need to design accurate AI/ML models so that their functioning is understandable to a non-technical audience.
Fairness: Assessing and mitigating biases according to pre-established criteria.	Training data need to be defined according to specific criteria such as demographic characteristics (e.g., women over 70 years old)

Filling the regulatory void

Regulators worldwide are proposing and adapting laws to manage **AI risks** and **accountability**

🌍 Global Approaches to AI Regulation:

📌 **United States:** Sector-specific oversight (FDA, state laws) adapting to AI in healthcare

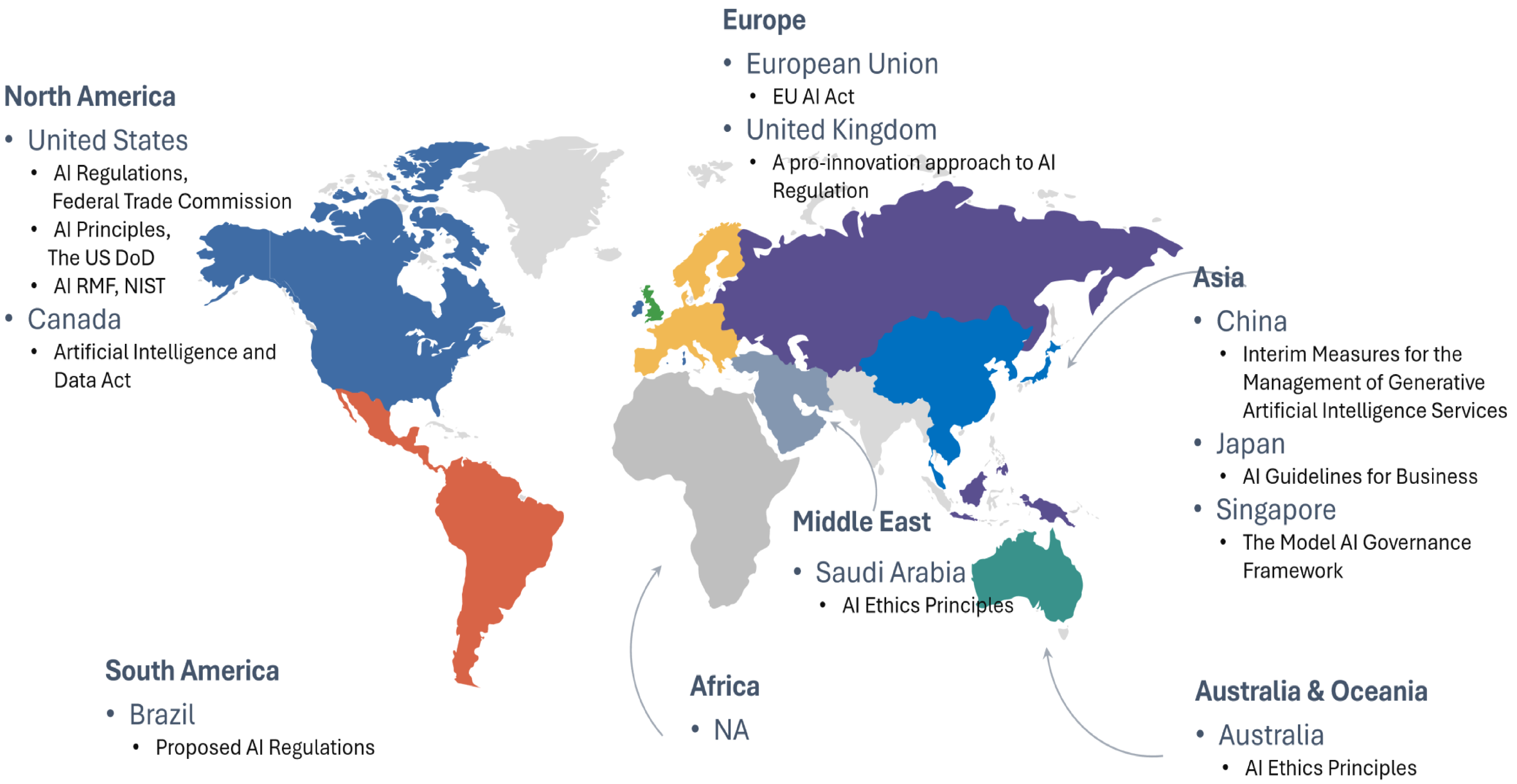
📌 **Australia:** Risk-based AI framework integrated with privacy and medical regulations

📌 **United Kingdom:** Pro-innovation approach aligning AI with existing legal structures

📌 **Japan:** Human-centric AI guidelines promoting ethical and responsible innovation

📌 **China:** Strict governance prioritizing security, ethics, and data control

📌 **European Union:** AI Act enforcing transparency, risk management, and compliance



North America
- United States
  - AI Regulations, Federal Trade Commission
  - AI Principles, The US DoD
  - AI RMF, NIST
- Canada
  - Artificial Intelligence and Data Act

Europe
- European Union
  - EU AI Act
- United Kingdom
  - A pro-innovation approach to AI Regulation

Asia
- China
  - Interim Measures for the Management of Generative Artificial Intelligence Services
- Japan
  - AI Guidelines for Business
- Singapore
  - The Model AI Governance Framework

Middle East
- Saudi Arabia
  - AI Ethics Principles

South America
- Brazil
  - Proposed AI Regulations

Africa
- NA

Australia & Oceania
- Australia
  - AI Ethics Principles

[Source: AI Watch: Global regulatory tracker (White and Case)]

The EU Approach

The EU seems to be leading with a comprehensive framework…

GOOD NEWS

The **EU AI Act** includes specific provisions for ***bias detection***, requiring that high-risk AI systems undergo rigorous testing and validation before their deployment in the EU marketplace.

- **First global AI framework** ensuring trustworthiness in Europe and beyond.
- **Effective**: August 1, 2024, with full applicability by August 2, 2026, and exceptions (prohibitions & AI literacy) from February 2025.
- **Risk-oriented regulation:** EU AI Act categorizes AI systems based on their risk level into four distinct categories: **unacceptable, high risk, limited risk, and minimal risk.**

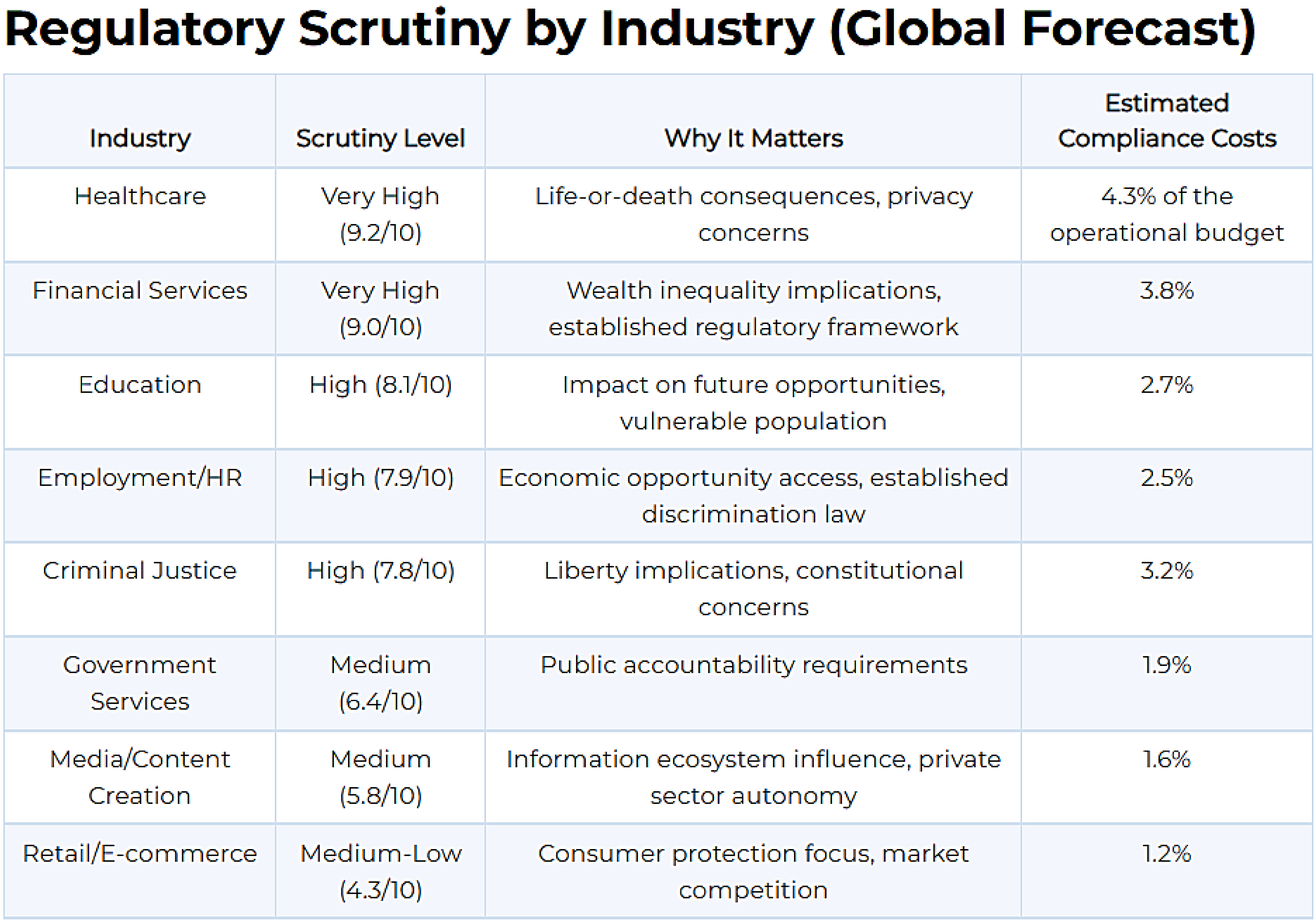
BAD NEWS

Compliance with the **EU AI Act (and other regulations)** may be difficult due to the rigorous regulatory requirements.

- **Compliance Costs:** Initial expenses for adherence to the Act.
- **Delays:** Increased administrative burden slows time to market.
- **Innovation Constraints:** Strict regulations could stifle R&D and innovation.
- **Uncertainty:** Practical applications may lead to ambiguous cases, necessitating court decisions to clarify interpretations.
- **GDPR & AI Act:** Increased complexity for multinational companies.

Regulatory Scrutinity

By 2030, industries like healthcare and finance will face the highest regulatory pressure to control AI bias.



Regulatory Scrutiny by Industry (Global Forecast)

Industry	Scrutiny Level	Why It Matters	Estimated Compliance Costs
Healthcare	Very High (9.2/10)	Life-or-death consequences, privacy concerns	4.3% of the operational budget
Financial Services	Very High (9.0/10)	Wealth inequality implications, established regulatory framework	3.8%
Education	High (8.1/10)	Impact on future opportunities, vulnerable population	2.7%
Employment/HR	High (7.9/10)	Economic opportunity access, established discrimination law	2.5%
Criminal Justice	High (7.8/10)	Liberty implications, constitutional concerns	3.2%
Government Services	Medium (6.4/10)	Public accountability requirements	1.9%
Media/Content Creation	Medium (5.8/10)	Information ecosystem influence, private sector autonomy	1.6%
Retail/E-commerce	Medium-Low (4.3/10)	Consumer protection focus, market competition	1.2%

https://www.allaboutai.com/resources/ai-statistics/ai-bias/#the-2030-ai-bias-index-exclusive-forecast

AI Implementation in Practice

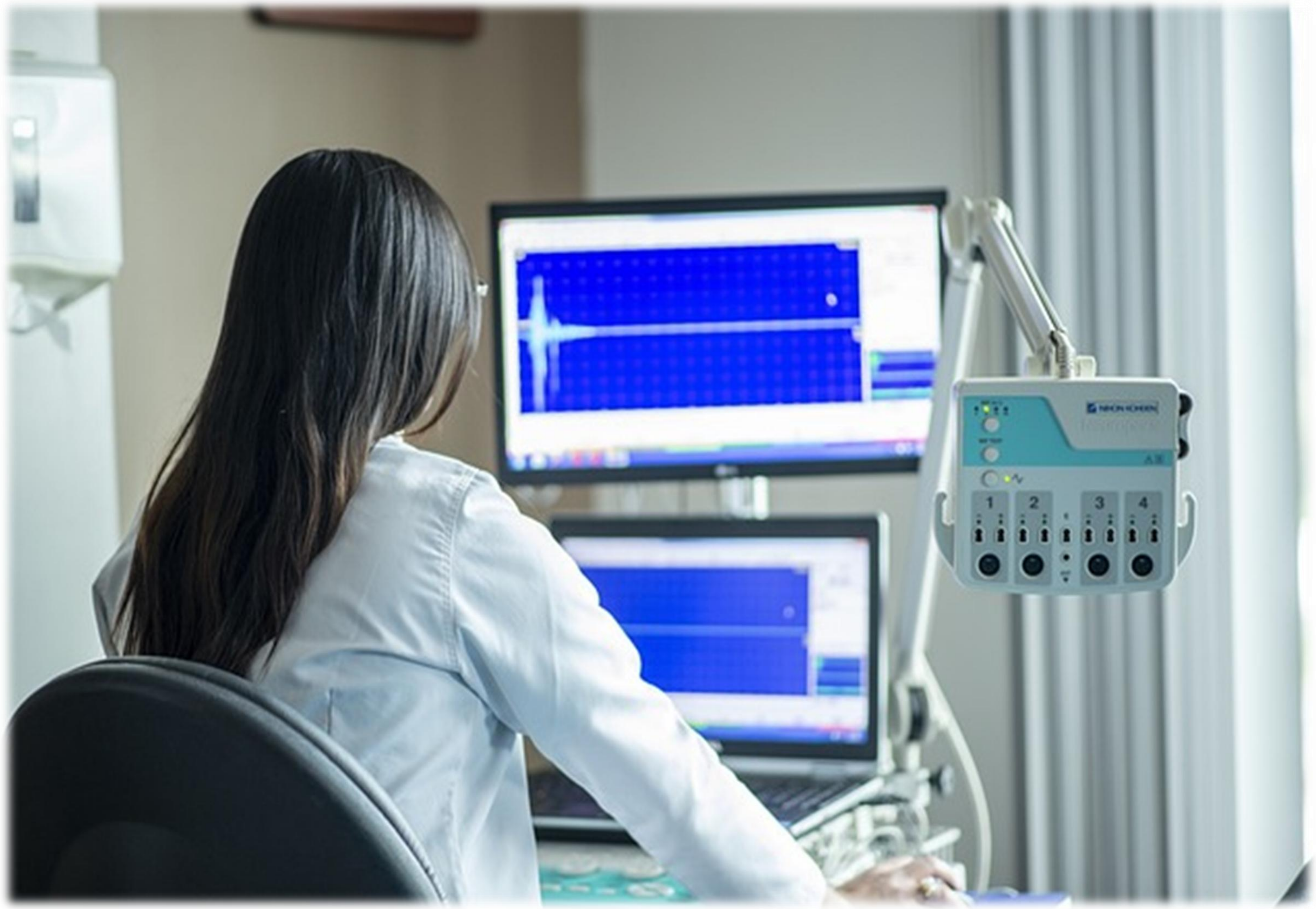
- Leading Spanish healthcare institution leveraging AI
- AI integration across multiple departments
- Focus on Radiology & Diagnostic Imaging

🎯 Goal: Boost efficiency & diagnostic accuracy

📖 This case study is available in the following chapter:

The Healing and Harmful Power of Data: Generative AI in Healthcare

📚 Trust in Generative Artificial Intelligence (Taylor & Francis)



Impact of AI on Patient Engagement & Clinical Outcomes

Reducing No-Shows with a Chatbot

– Chatbot reminders haven't eliminated no-shows entirely (e.g., accidents, emergencies) but have significantly **cut absenteeism**.

– Empowers patients: "*I cannot come—please give me another appointment*," boosting rescheduling and reducing wasted slots.

– Better resource allocation → **more patients seen on time.**

Enhancing Cancer Detection with Computer-Aided Detection (CAD) + (Fully Automated Detection Methods (FFDM)

–AI-assisted mammography **improved detection rates** by up to 16–20%.

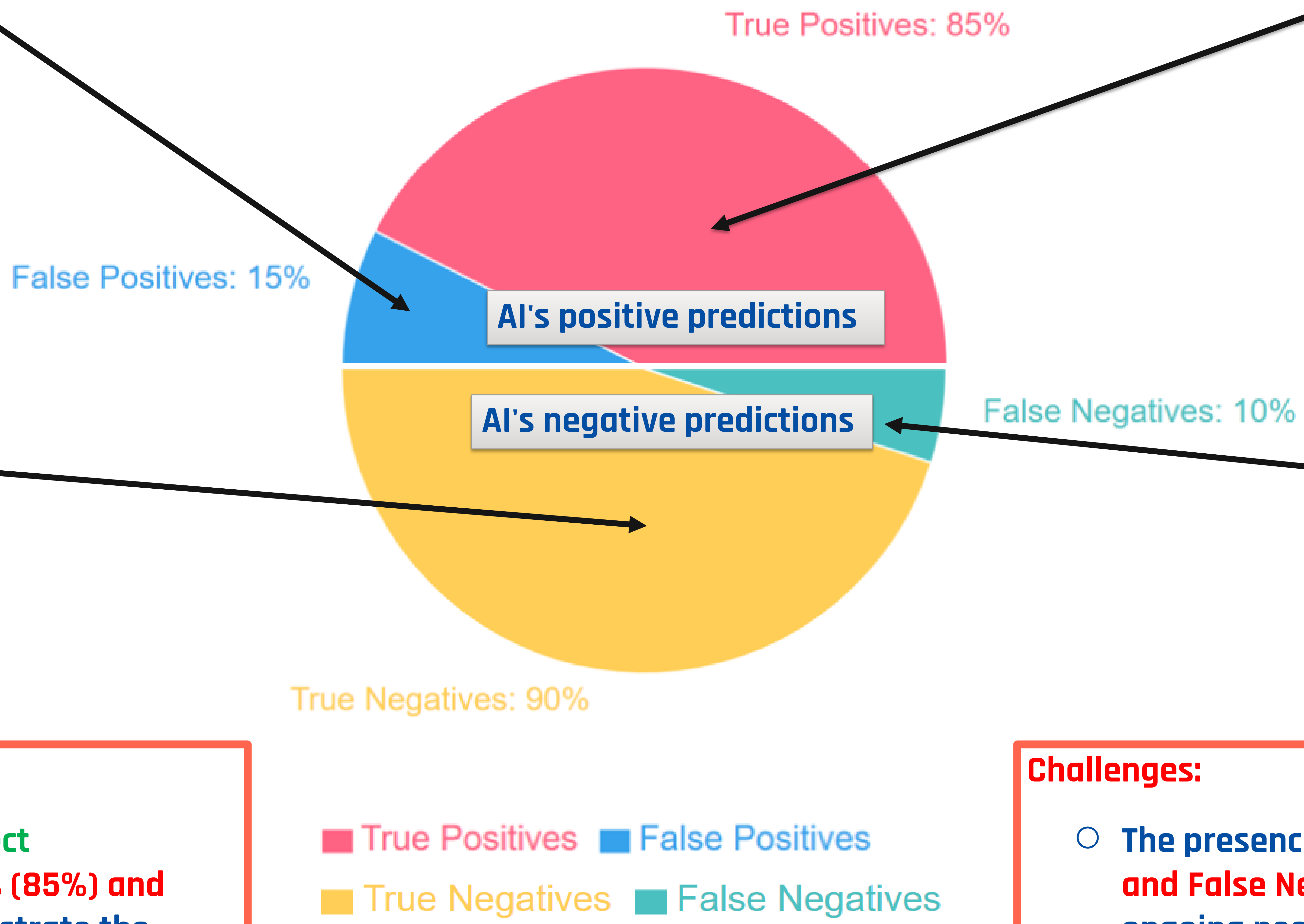
– A **CAD-supported single-reader workflow** proved as **effective** as traditional double reading with arbitration.

– Automation of routine tasks **freed radiologists' time** and **increased sensitivity for calcifications** and small invasive cancers.

Specific AI Application: Cancer screening

- Indicates **room for improvement in reducing false alarms** (e.g., lower sensitivity for certain types of lesions)
- May lead to **increased workload**, **unnecessary stress** and **additional testing** for patients

- Correctly **identifies non-cancer cases**
- **Reduces unnecessary anxiety and further testing** for healthy individuals
- Helps in **efficient resource allocation** in healthcare systems



- **High accuracy** and **sensitivity** in detecting actual cancer cases (e.g., enhanced identification of calcifications and small invasive cancers)
- Contributes significantly to **early detection** and potential life-saving diagnoses

- Most critical area for improvement due to **potential missed cancer diagnoses**
- May lead to **increased workload**, **unnecessary stress** and **additional testing** for patients

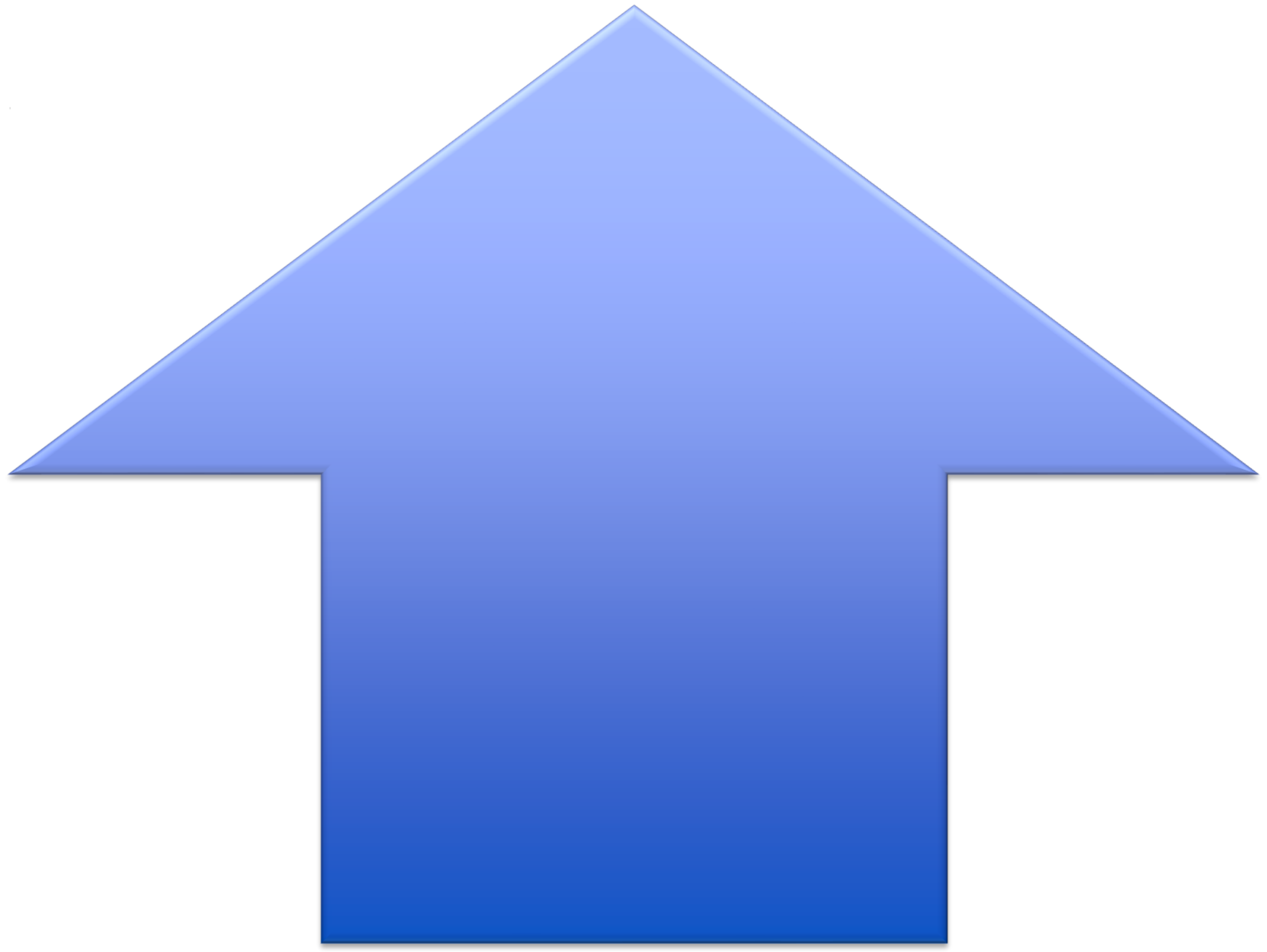
Benefits:

- **The high percentages of correct identifications (True Positives (85%) and True Negatives (90%)) demonstrate the overall effectiveness of the AI system**

Challenges:

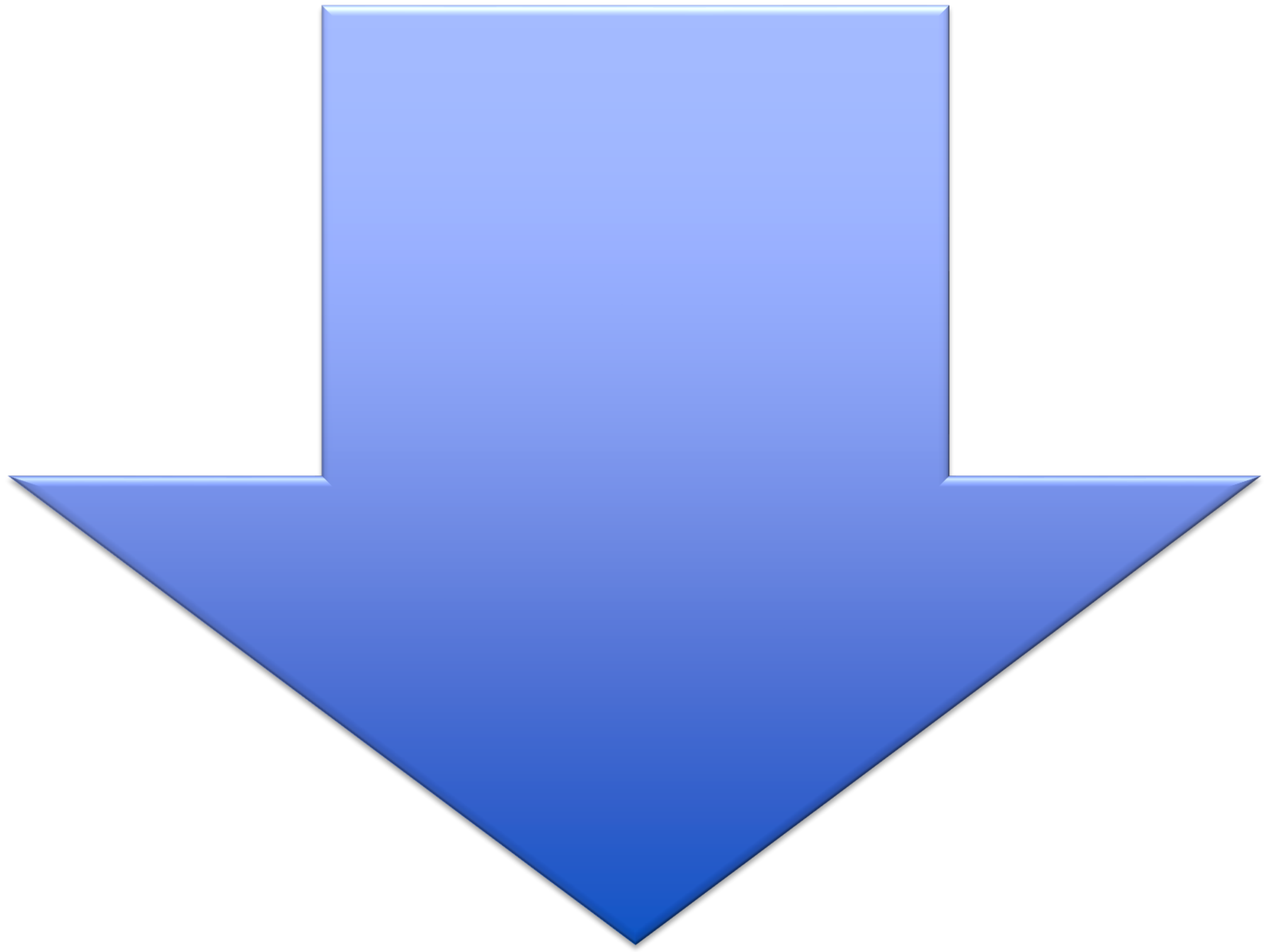
- **The presence of errors (False Positives (15%) and False Negatives (10%)) highlights the ongoing need for human oversight and system refinement, especially for hard-to-detect cases**

Under what conditions can AI enhance or hinder healthcare progress?



AI can improve healthcare when:

- Models are trained on **diverse, representative data sets** to avoid under- or over-serving any patient group.
- Development and deployment processes include rigorous **bias audits**, **transparency reports**, and **stakeholder feedback loops**.
- Clinicians and data scientists collaborate closely, **combining human judgment with algorithmic insights** to catch and correct systemic skew.
- **Ongoing monitoring** tracks performance across demographics, with mechanisms to recalibrate models when disparities emerge.



AI can hinder progress when:

•**Training data reflect historical biases**, leading to uneven accuracy across patient populations.

•Implementation leads to **over-reliance on technology**, potentially increasing false positives or recall rates.

•**Data security and privacy gaps can expose sensitive attributes**, enabling biased inferences.

•There's **insufficient transparency** around how algorithms make decisions, preventing clinicians and patients from identifying biased outputs.

•Continuous **evaluation** and **feedback** channels are **absent**, so biased predictions go unchecked and become entrenched in care pathways.

A Scientific Simulation Based on Gender-Differentiated Outcomes in Cardiac Diagnostics

Case Study: CardioLens AI Diagnostic System

- **System Architecture**
  - **CardioLens AI** is a machine learning-based Clinical Decision Support System (CDSS) that evaluates **cardiac risk** using:
    - **Core engine:** Gradient boosting algorithm trained on 1.2M patient records
      - Advanced machine learning technique, iteratively combines multiple decision trees to create a **powerful predictive model**
    - **Inputs:** 150 clinical features to predict 8 cardiac conditions
      - Processes a **wide range of patient data** (e.g., vital signs, lab results, demographics, symptoms) to assess **multiple heart problems**.
    - **NLP module:** Natural language processing of physician notes
      - Analyzes and interprets **unstructured text** in doctors' notes to extract relevant medical information for the AI system
    - **Integration:** hooks directly into **electronic health-record** systems for real-time data access
      - Allows the AI to instantly retrieve and **analyze up-to-date patient information** from hospital databases
    - **Explainability:** Interpretability layer using (SHapley Additive exPlanations) values
      - Provides insights into how the AI makes decisions, helping doctors **understand** and **trust** the system's recommendations
    - **Rollout:** live in 6 leading university hospitals
      - Implemented in high-level academic medical centers for **real-world testing** and **application**.

The Problem

- **Performance Metrics Analysis:**
  - **Overall accuracy: 89% of cardiac risks identified (compared to expert cardiologist panel)**
  - **Male patient accuracy: 95% (n=3,240 patients)**
    - For **men**, the system is **highly accurate**, with only 5% error rate in a large sample.
  - **Female patient accuracy: 76% (n=2,890 patients)**
    - For **wome**n, the **accuracy drops** significantly, with a 24% error rate in a similarly large sample.
  - **Statistical significance: $p<0.001$**
    - The gender-based **accuracy difference** is not due to chance and represents a **real disparity**.

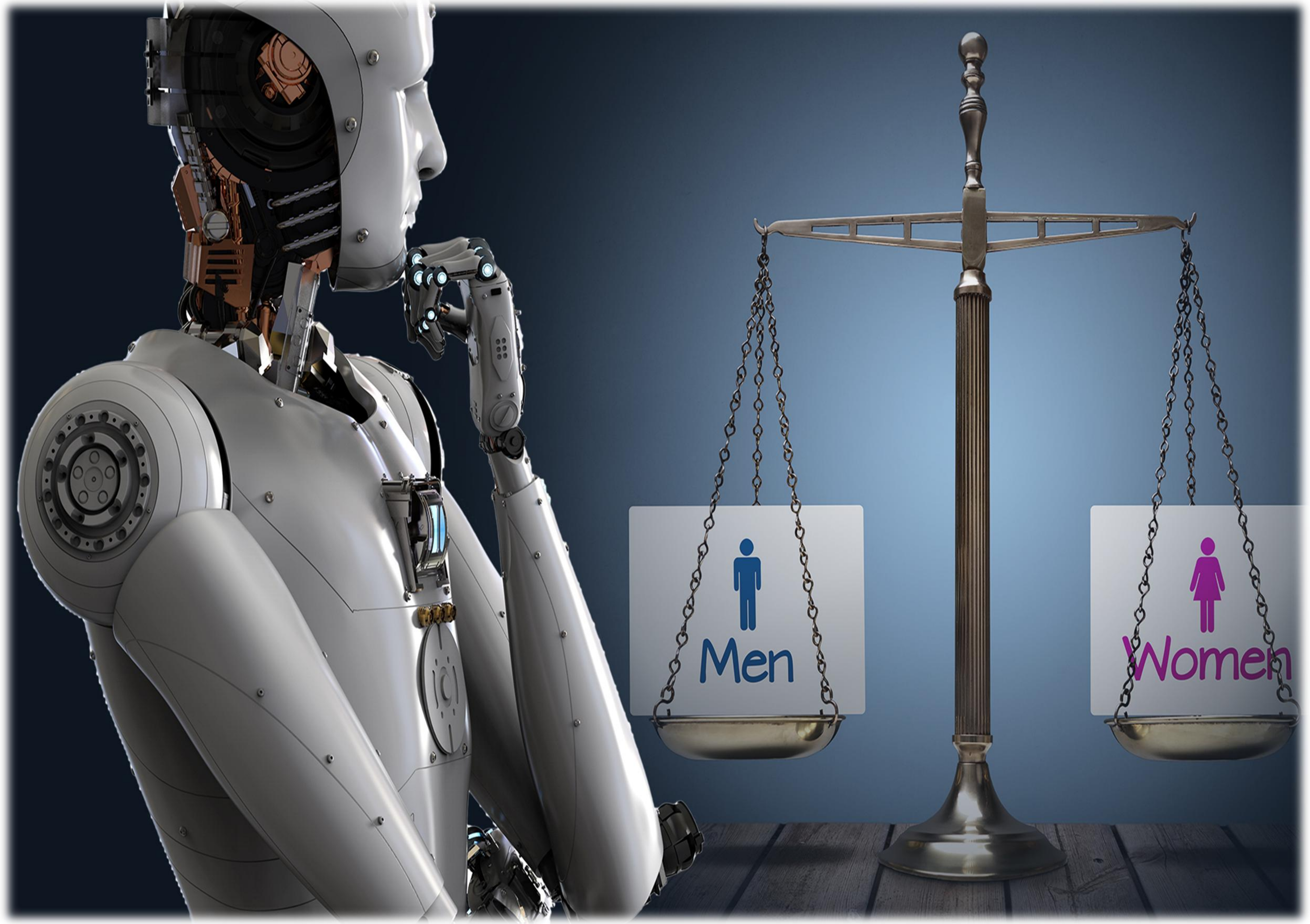
- **Gender-Differentiated Diagnostic Patterns**

Acute coronary syndrome detection:

  - **Sensitivity: 72% accurate in women vs. 94% accucare in men**
    - The system is much better at identifying heart attacks and similar conditions in men than in women.
  - **False negative rate for women with “atypical” symptoms: 43%**
    - Nearly **half of women** presenting with non-standard heart attack symptoms are **incorrectly classified** as low-risk.
  - **Algorithm confidence levels average 12% lower for female patients**
    - The AI is less certain about its diagnoses for women, indicating **potential gaps** in its understanding of **female cardiac issues.**

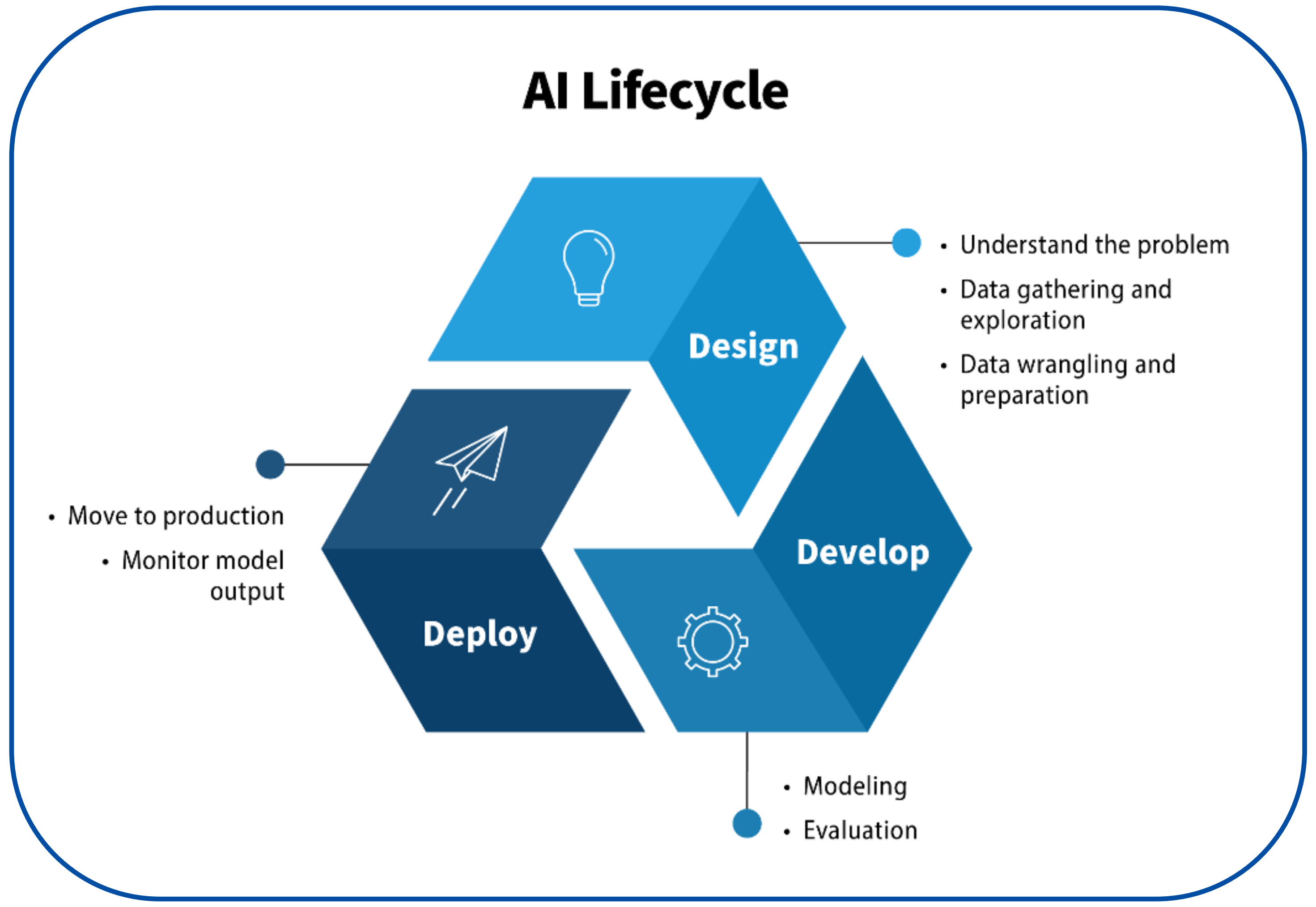
Impact on Patient Outcome:

- **Misdiagnosed women: 2.7x higher 30-day mortality rate**
  - Women **wrongly assessed** by the AI are almost three times more likely to die within a month compared to correctly diagnosed patients.
- **Women face 4.2 hours longer treatment delays on average**
  - Female patients wait **over 4 hours longer** for **appropriate cardiac care**, potentially due to AI misdiagnosis.
- **68% more women return within 72 hours**
  - Many more **women come back** to the hospital within **three days**, suggesting **initial misdiagnosis** or **inadequate treatment.**



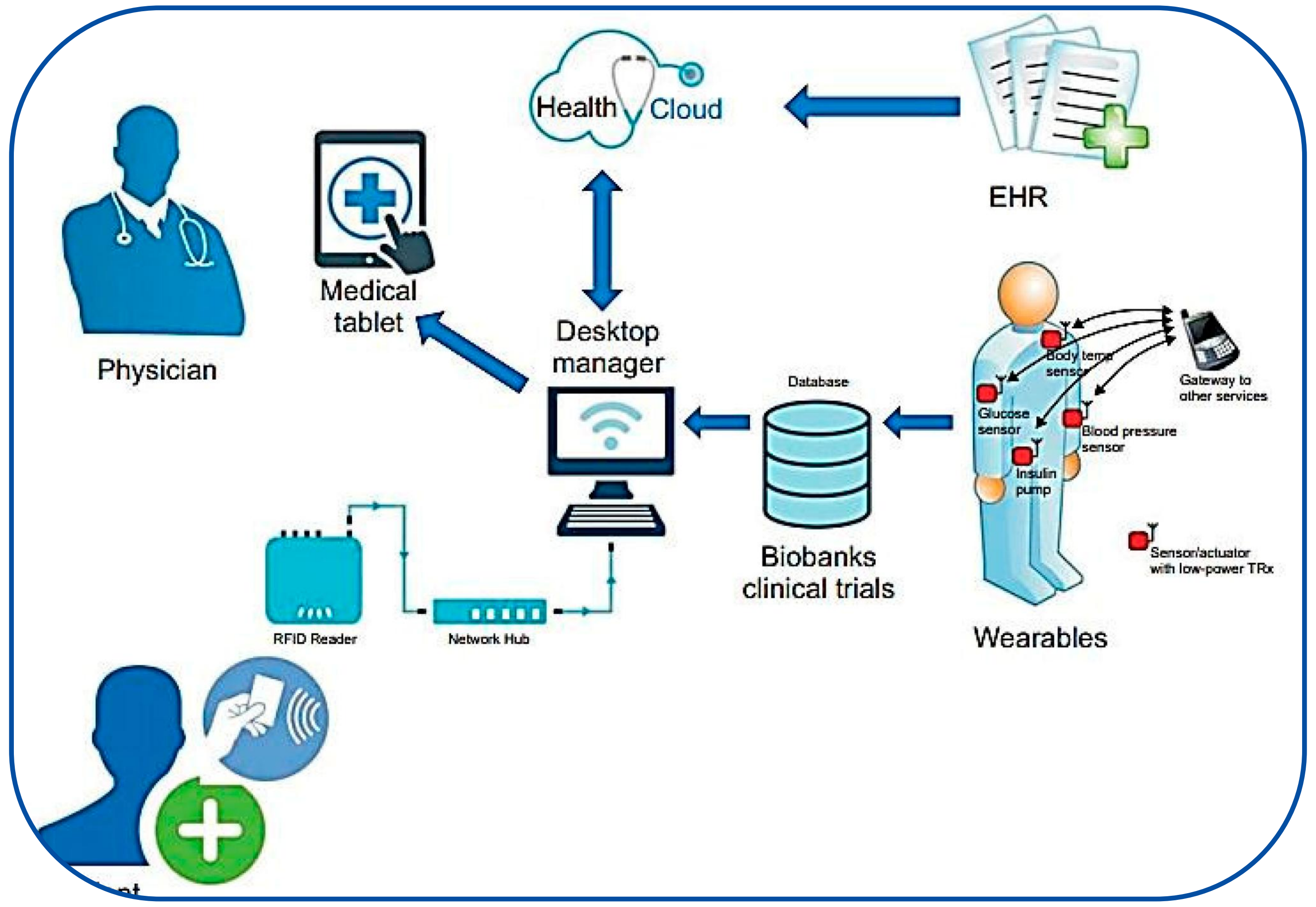
Hypothesis: The system exhibits algorithmic bias through systematically undervaluing female-predominant symptom presentations.

Relevance to Subsequent Analysis:



Performance disparities highlight potential biases in the AI lifecycle

Outcome differences emphasize the clinical impact of these biases

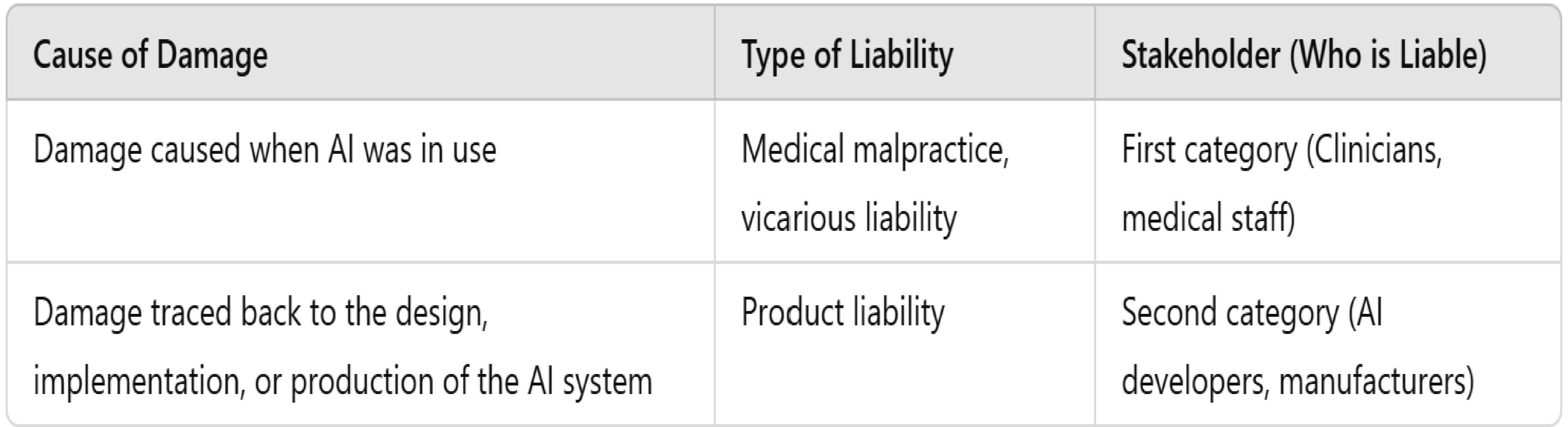


System architecture details provide insight into potential bias sources

We'll use a forward-backward analysis to examine the CardioLens AI system. First, we'll look at outcomes (forward), then investigate causes (backward).

Forward Analysis (CardioLens)

- **Key Info:**
  - **AI System (CardioLens):**
    - Gradient boosting algorithm, 1.2M patient records
    - NLP for physician notes, EHR integration
    - SHAP values for explainability
  - **Gender Disparities:**
    - Women: 2.7x higher 30-day mortality if misdiagnosed
    - 4.2 hours longer treatment delays for women
    - 68% more women return within 72 hours



Cause of Damage	Type of Liability	Stakeholder (Who is Liable)
Damage caused when AI was in use	Medical malpractice, vicarious liability	First category (Clinicians, medical staff)
Damage traced back to the design, implementation, or production of the AI system	Product liability	Second category (AI developers, manufacturers)

Point of Discussion:

- Delineation of **responsibilities** between human operators and AI system in decision-making processes
- Comprehensive assessment of **cause of damage**
- **Stakeholder identification** (individuals involved in AI use and development)

Liability Resolution Framework for CardioLens AI System

Damage Caused When AI Was in Use

- **Cause of Damage: Direct consequences of AI system during diagnostic process evidenced by:**
  - 2.7x higher 30-day mortality for misdiagnosed women
  - 4.2 hours longer treatment delays
  - 68% more women returning within 72 hours
- **Liability Type**: **Medical malpractice** and **vicarious liability**
- **Stakeholders: Clinicians and medical** staff who directly use the AI system
- **Key Responsibilities:**
  - Ensure accurate **patient care**
  - Critically **evaluate AI recommendation**s
  - Maintain **independent clinical judgment**
  - Provide **comprehensive patient monitoring**

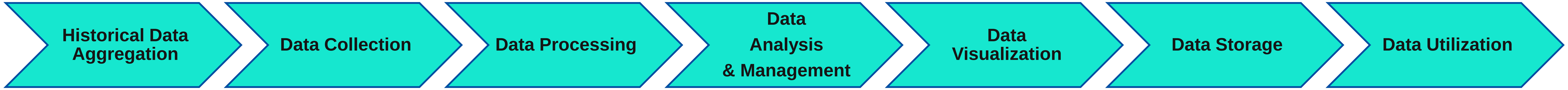
Damage Traced to AI System Design

•**Cause of Damage:** Directly related to the **structural and design-level issues** in the AI system, potentially:

  - NLP module misinterpreting gender-specific medical language
  - SHAP values revealing algorithmic bias
- **Liability Type: Product Liability**
- **Stakeholders:** AI **Developers** and **Manufacturers** (CardioLens development team responsible for system design)
- **Key Responsibilities:**
  - Develop **unbiased machine learning algorithms**
  - Ensure **representative training data**
  - Implement **robust NLP interpretation**
  - Create **transparent explainability** mechanisms and continuously monitor and correct biases

Backward Analysis

- **Key Info:**
  - **AI System (CardioLens):**
    - Gradient boosting algorithm, 1.2M patient records
    - NLP for physician notes, EHR integration
    - SHAP values for explainability
  - **Gender Disparities:**
    - Women: 2.7x higher 30-day mortality if misdiagnosed
    - 4.2 hours longer treatment delays for women
    - 68% more women return within 72 hours



Historical Data Aggregation

- Source of bias (Systemic Bias): Societal prejudices
- Example: Studies on HIV and Autism predominantly contain data on male patients

Data Collection

- Source of bias (Selection Bias): Personal biases of data collectors
- Example: Unconscious selectivity against women in treatment data gathering

Data Processing

- Source of bias (Algorithmic Bias): Lack of context understanding
- Example: Misinterpretation of weight measurements without considering gender differences

Data Analysis & Management

- Source of bias (Evaluation Bias): Statistical bias
- Example: Analysis of electronic health records (EHRs) lacking information on women

Data Visualization

- Source of bias (Reporting Bias): Unconscious biases in data labeling
- Example: Mislabeling critical indicators due to limited exposure to female-specific symptoms

Data Storage

- Source of bias (Confirmation Bias): Insecure storage of biased data
- Example: Unencrypted storage leading to potential exploitation by cyber attackers

Data Utilization

- Source of Bias (Generative Bias): Biased decision-making in healthcare applications
- Example: AI misdiagnosing heart attacks in women due to bias

Point of Discussion:

- Map and address potential **biases across the entire AI data lifecycle**
- Detect **discrepancies** in **data representation, collection, and interpretation** between male and female patients

Bias Source Resolution Framework for CardioLens AI System

- **Historical Data Aggregation**
  - **Bias Source:** Systemic bias in data collection
  - **Evidence:** 1.2M patient records potentially skewed
  - **Impact:** Unequal representation of patient demographics
- **Data Collection**
  - **Bias Source:** Unconscious selectivity in data gathering
  - **Evidence:** Bias against women in treatment data collection
  - **Impact:** Misrepresentation of female patient characteristics
- **Data Processing**
  - **Bias Source:** Algorithmic misinterpretation
  - **Evidence:** Medical parameters without gender context
  - **Impact:** 95% accuracy for male patients vs. 76% for female patients
- **Data Analysis**
  - **Bias Source:** Statistical bias
  - **Evidence:** EHRs lacking information on women
  - **Impact:** 12% lower diagnostic confidence for women
- **Data Visualization**
  - **Bias Source:** Reporting bias
  - **Evidence:** Mislabeling female-specific symptoms
  - **Impact:** 43% false negative rate for women
- **Data Storage & Utilization**
  - **Bias Source:** Insecure data handling
  - **Evidence:** Potential data exploitation
  - **Impact:** Compromised data integrity
- **Data Utilization**
  - **Bias Source:** Biased decision-making
  - **Example**: AI misdiagnosing heart attacks in women
  - **Impact:** 2.7x higher 30-day mortality for misdiagnosed women

Key Recommendations

For The Forward Analysis

- Implementation of **clear policies** for AI system use and clinician training
- **Risk management strategies** for potential **vicarious liability** scenarios
- Protocols for handling cases where **damage may be attributed to both AI and human factors**
- **Documentation** of clinical **decision-making** process when using or overriding AI recommendations
- Balancing **AI reliance** with **independent clinical judgment**, especially for **atypical presentations**

For the Backward Analysis:

- Create a **multi-stage bias correction protocol** identifying potential bias sources across the AI data lifecycle.
- **Prioritize gender-inclusive data collection** and **algorithmic refinement** to reduce diagnostic inequities and improve overall patient outcomes.
- Develop targeted strategies to **improve diagnostic accuracy** and r**epresentation** for female patients (e.g., Adaptive learning mechanisms)
- Regular **algorithmic audits**

Common: Awareness of gender-specific symptoms and AI system limitations

Conclusions

Can we create a *neutral* AI? Probably not, but….

- **Progress requires collaboration across:**
  - Data scientists
  - Healthcare providers
  - Regulators
  - Legal experts
- **Ongoing assessment** of when automated decision-making is appropriate
- Combined **AI-human approach** can reduce bias (**Human in the Loop**)
- Turn AI from challenge to **opportunity**

Any questions?

Email:
amarotta@mit.edu

References & Things to Read:

- Marotta, A. (2025). **Trust in generative AI within healthcare: Healing and harmful powers of data**. In J. Paliszkiewicz, I. Dąbrowski, & L. Halawi (Eds.), Trust in generative artificial intelligence: Human-robot interaction and ethical considerations. Routledge, Taylor and Francis. Available at **https://www.taylorfrancis.com/chapters/edit/10.4324/9781003586937-15/healing-harmful-powers-data-angelica-marotta**
- Marotta, A. (2022) **When AI Is Wrong: Addressing Liability Challenges in Women's Healthcare**, Journal of Computer Information Systems, DOI: 10.1080/08874417.2022.2089773 **Available at https://www.tandfonline.com/doi/full/10.1080/08874417.2022.2089773**