CAIR-Nepal

Shaping the Future with AI

CAIR-Nepal
Shaping the Future with AI

Emergence of Biases in AI

Abhash Shrestha

02, 01, 2025

University of Southampton — MSc Computer Science

CAIR-Nepal (Shaping the Future with AI) — Research

01000010 BINARY DIGITS — Co-founder

Outline

1. Introduction - Bias in AI: Why We Should Care
2. Symbolic AI
3. Sub-Symbolic AI
4. Understanding AI Biases
5. Standard Approaches to Mitigate AI Bias
6. Combining Symbolic & Sub symbolic AI

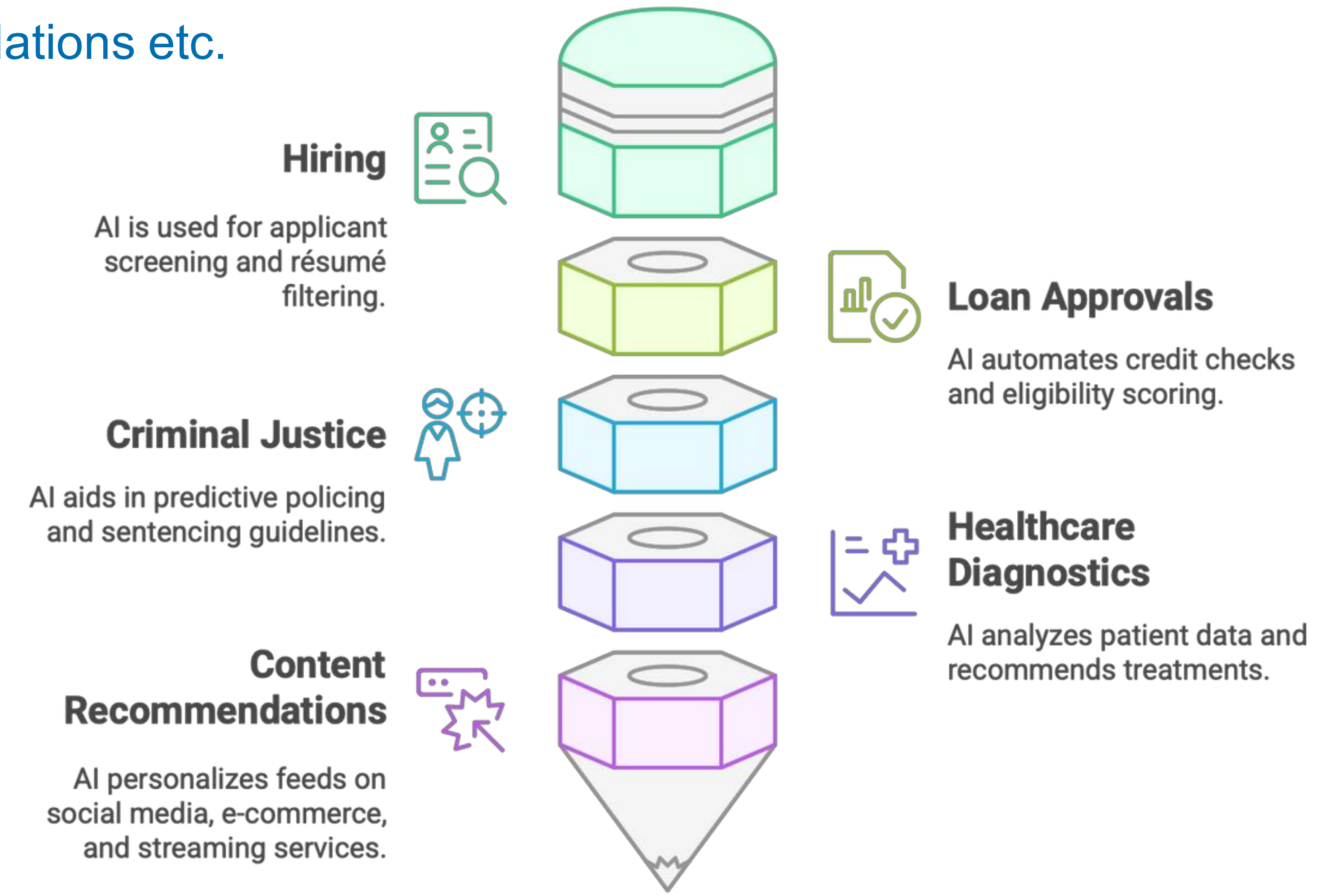
Why Deal with AI biases?

Is it even a problem worth solving?

Imagine applying to your dream job and never getting an interview because the AI screening résumés learned to favor certain demographics. Real incidents like Amazon’s hiring tool highlight the human impact of AI bias and why addressing it is so urgent.

1. Introduction - Bias in AI: Why We Should Care

- **Prevalence of AI**
  - increasing integration into decision-making processes
    - hiring
    - loan approvals
    - criminal justice
    - healthcare diagnostics
    - content recommendations etc.



IEEE CAIR-Nepal Shaping the Future with AI

Consequences

Discrimination in Hiring

Algorithms favor certain demographics, leading to unfair job application outcomes.

Societal Disharmony

Marginalized communities lose trust in AI, deepening societal divides.

Low Trust in Technology

Biased outcomes erode public confidence in AI systems.

Unfair Policing

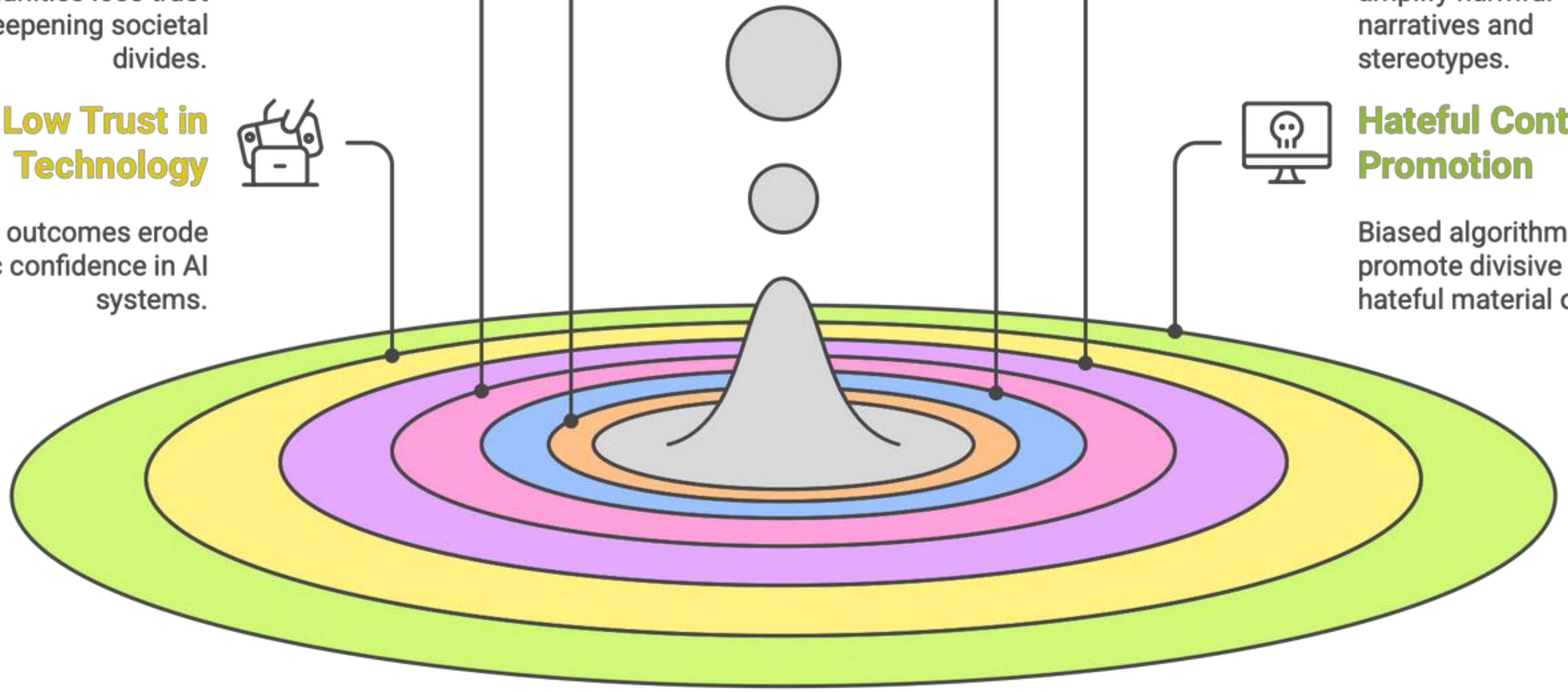
Biased data results in disproportionate surveillance and harsher punishments.

Perpetuation of Stereotypes

Automated systems amplify harmful narratives and stereotypes.

Hateful Content Promotion

Biased algorithms promote divisive or hateful material online.



- **Real-World Consequences**
  - Biased AI can lead to unfair treatment
    - gender or racial discrimination in job applications
    - unfair policing and sentencing

- **Societal Impact**:
  - Societal Disharmony
  - Perpetuation of Stereotypes
  - Low trust in Technology
  - Hateful Content in AI and internet

- **Real-World Example:**
  - Amazon’s Hiring Tool
    - Internal AI recruitment system - penalized résumés containing the word “women’s.”
    - Shows how historical data (dominated by men in tech roles) can inadvertently encode discrimination

2. Symbolic AI - Good Old Fashioned AI

- Symbolic AI represents knowledge through symbols - words, logical rules, and explicit ontologies.
- Historically dominated early AI research (1950s–1980s).

Key Characteristics

- Knowledge Representation
  - Information is encoded in a structured, human-readable way
- if-then rules
- decision trees
- first-order logic

Reasoning

- Symbolic systems use logical inference engines to draw conclusions from known facts.

Example

- IBM's chess-playing system that used **symbolic search trees and rule-based evaluation functions** to defeat Garry Kasparov.

Examples & Applications

Expert Systems (e.g., MYCIN for medical diagnosis): If symptom X and symptom Y, then consider disease Z.

Knowledge Graphs in semantic web and enterprise settings (e.g., storing relationships: “Product X is made by Company Y”).

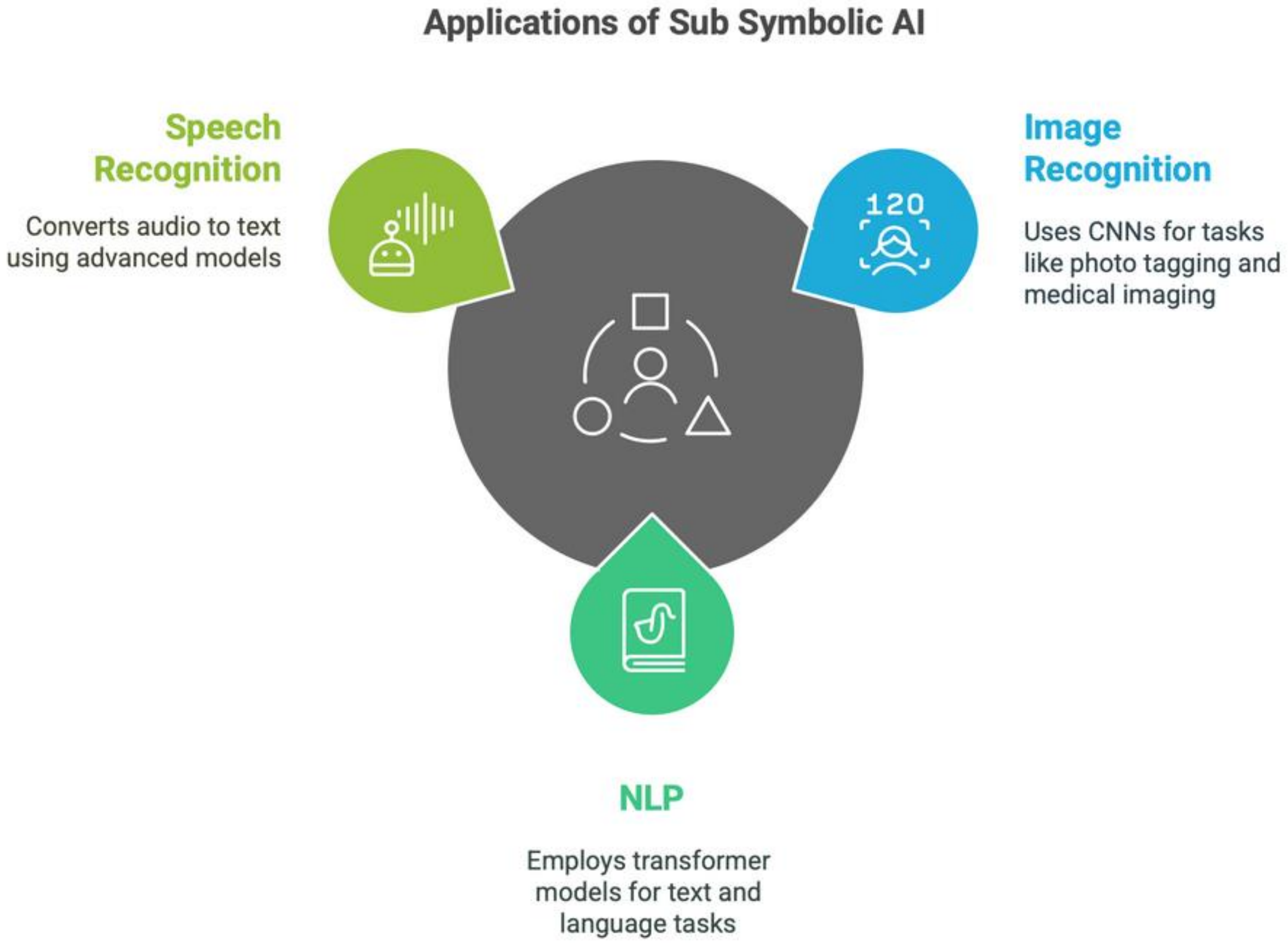
Why use them?
Easy interpretation : You can trace how the system arrived at a conclusion via logical steps.
Easy Update: Easy to identify and modify rules if a certain rule is found to be biased or incorrect.
Domain Knowledge: Allows easy incorporation of formal domain knowledge (e.g., legal constraints, medical guidelines).

3. Subsymbolic AI

- Most popular these days
- Machine Learningn, Deep Learning
- focuses on learning patterns from data instead of explicitly programmed rules.

Examples & Applications

- Image Recognition: Convolutional Neural Networks (CNNs) used in photo tagging, medical imaging diagnostics.
- Natural Language Processing (NLP): Transformer-based models (e.g., GPT, BERT) for text generation, language translation, chatbots.
- Speech Recognition: Recurrent or transformer-based models that convert audio waveforms to text.



Weaknesses

- Black-Box Nature
  Difficult to interpret or explain decisions
  e.g: why did the model classify this person as high risk?

- Data Dependency: The model's accuracy and fairness heavily depend on the training data's quality and representativeness.
- Emergence of Bias - herein lies it's susceptibility to biases

4. Understanding AI Bias

- Deviation from a “true” or “fair” outcome, leading to disproportionate negative impacts on specific groups or individuals.

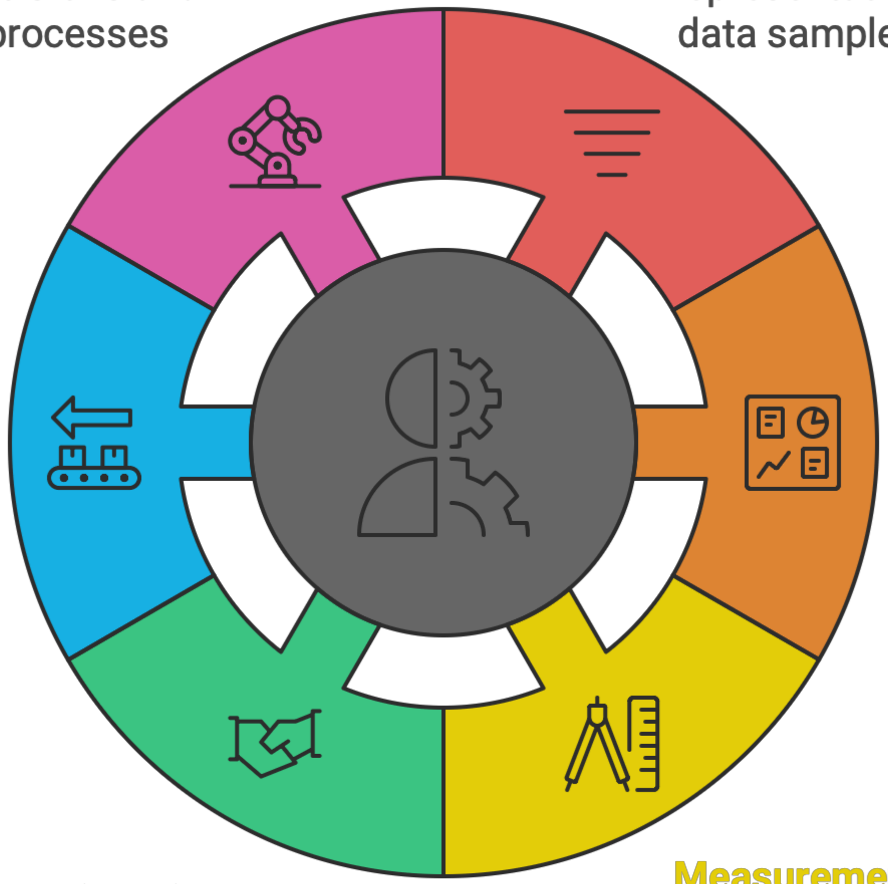
- Why Does Bias Occur in AI?

- Historical & Societal Patterns: AI learns from data reflecting past and present inequalities (e.g., fewer women in tech leads to biased hiring models).
- Data Collection & Curation: Datasets might overrepresent certain demographics and underrepresent others, reinforcing stereotypes.

Some Consequences

- **Reinforcement of Stereotypes**: Language models generating text that stereotypically portrays women in caretaker roles and men in leadership roles.
- **Discrimination**: Banks offering unfair loan terms to minority applicants.

Some common types of Biases



Algorithmic Bias

Biases from algorithmic decisions and processes

Sampling Bias

Biases arising from non-representative data samples

Representation Bias

Biases due to inadequate or skewed data representation

Measurement Bias

Biases from inaccurate data measurement tools

Interaction Bias

Biases from user interactions affecting data outcomes

Generative Bias

Biases from model assumptions during data generation

5. Standard Approaches to Mitigate AI Bias

Data-Centric Approaches

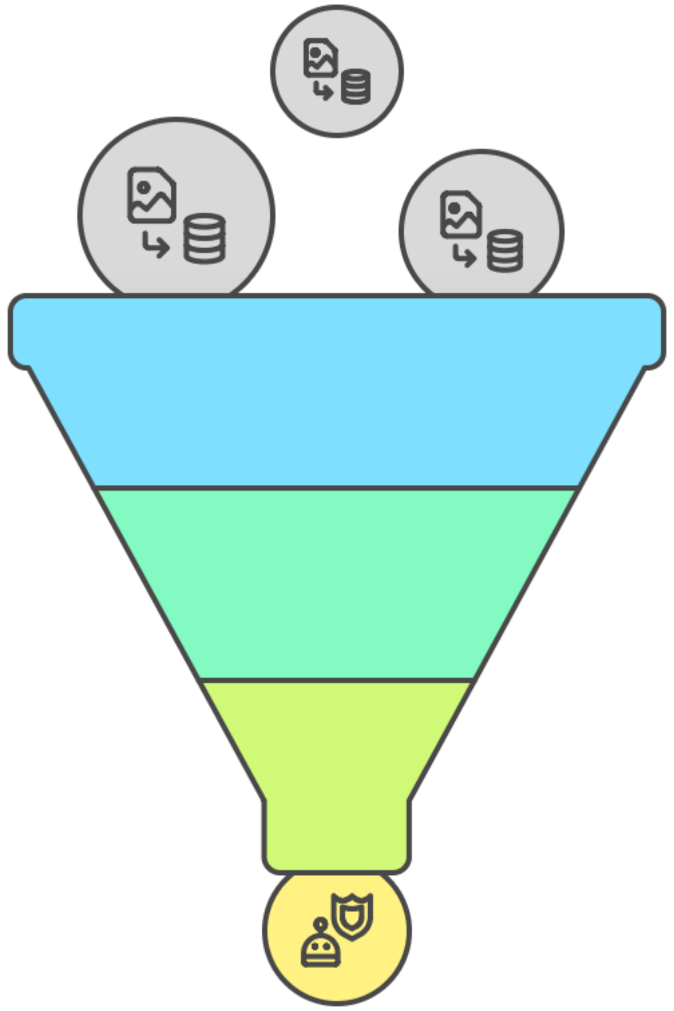
Data Collection (Preprocessing):
- Use balanced datasets or oversampling techniques for underrepresented groups.
- Perform de-biasing transformations (eg: removing or anonymizing sensitive attributes).

Algorithmic Approach (In-processing)

Fairness-Constrained Learning:
- Incorporate fairness metrics (e.g., demographic parity, equalized odds) into the training objective.
- Example: Adjust predictions to ensure false positive rates are equitable across demographic groups.

Auditing & Explainability (Post-processing)
- Interpretability Tools: LIME, SHAP, or integrated gradients to highlight which features are most influential in a decision.
- Fairness Dashboards & Toolkits: IBM AI Fairness 360, Microsoft Fairlearn, and Google's Responsible AI toolkit offer metrics and visualizations to diagnose biases.

Bias Mitigation Process



Data Collection -Pre processing

Preparing data by balancing and de-biasing

Algorithmic Approach - In processing

Applying fairness metrics during training

Auditing & Explainability - Post processing

Using tools to interpret and visualize outcomes

6. Combining Symbolic & Subsymbolic AI

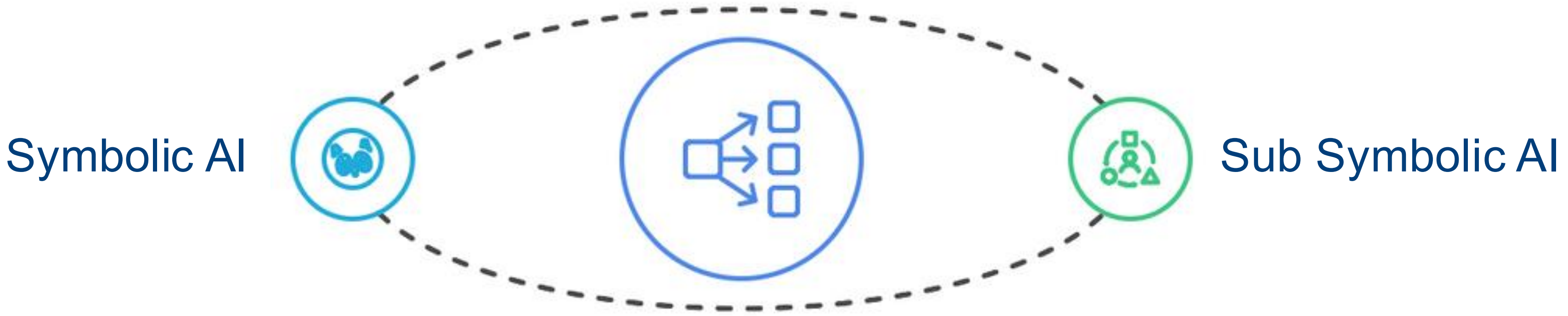
A Hybrid Approach

Complimentary Strengths

- Symbolic AI => explicit, logical constraints about fairness, ethical principles, or legal compliance.
- Sub-symbolic AI => excels at pattern discovery and handling high-dimensional data.

Mitigating Bias

- Rule-Based Fairness Constraints
  - Use Symbolic logic
  - define rules
  - eg: “candidates with equivalent qualifications must have comparable scores regardless of demographic group”
- Data-Driven Insight
  - Neural networks can detect complex, non-obvious relationships
  - Guided by symbolic logic to avoid discrimination



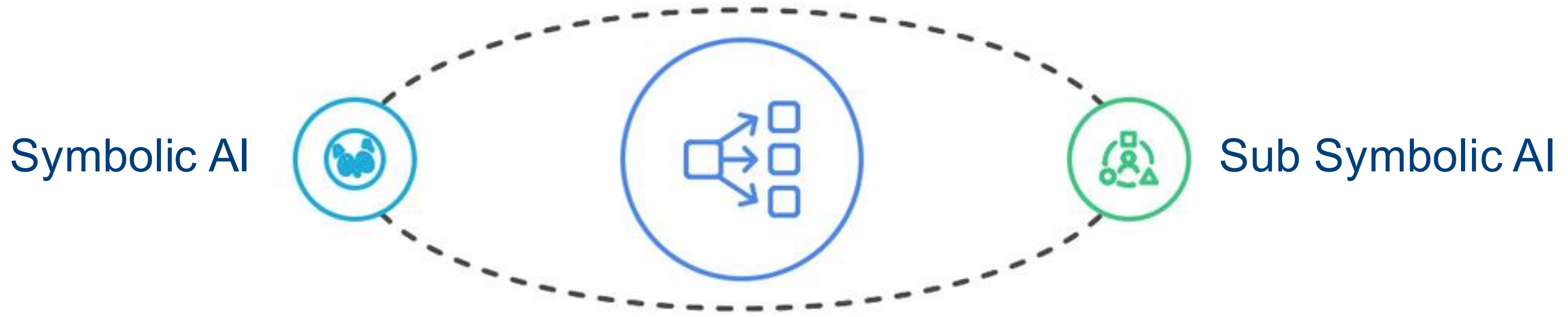
]

A Hybrid Approach

Interpretability & Accountability

- Output Explanations:
  - Symbolic layer - store relations
  - decisions can be traced back to them

- Error Checking and Preventing Discrimination:
  - symbolic rules can act as a safety net
  - If sub-symbolic AI (neural model) - output violates a fairness constraints
  - system can adjust or override the decision

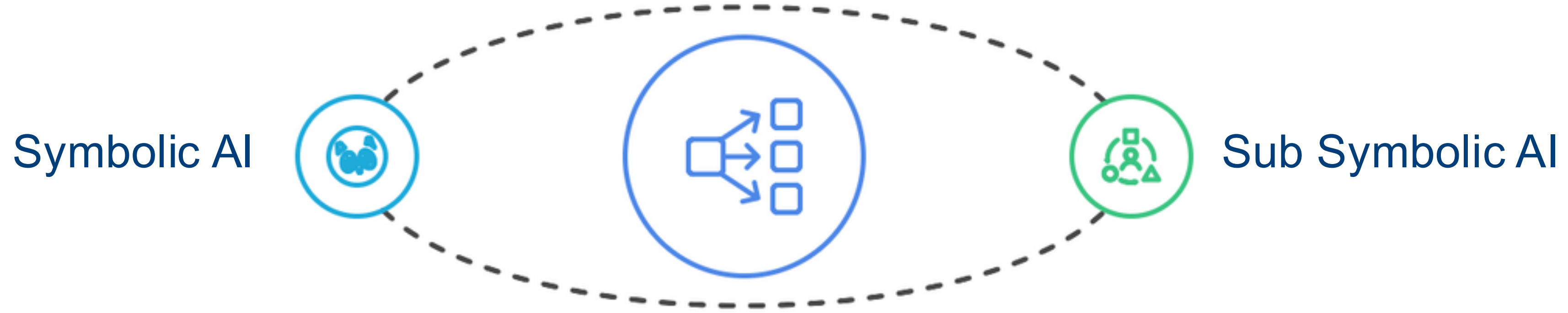


IEEE CAIR-Nepal Shaping the Future with AI

A Hybrid Approach

Challenges & Limitations of the Hybrid Approach

- Increased System Complexity
  - Merging symbolic and subsymbolic components can lead to more intricate architectures that are harder to maintain and debug.
- Ongoing Rule Management
  - Symbolic rules must be frequently updated as definitions of fairness or regulations change (e.g., new protected categories).
- Performance vs. Fairness Trade-Off
  - Imposing fairness constraints can sometimes reduce raw accuracy, creating tension between performance and social responsibility.



Thank you!