



# Breaking the Mirror: Self-Reflection as an Approach to Mitigating Inherited Biases in AI Models

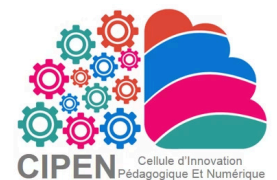
We'll discover how artificial intelligence with **self-reflection mechanisms** can transform from **perpetuating stereotypes** to actively **promoting equality**.

**Dr. Jordi Linares** (jorlipel@upv.es)

**Universitat Politècnica de València (SPAIN)**



Co-funded by the  
European Union

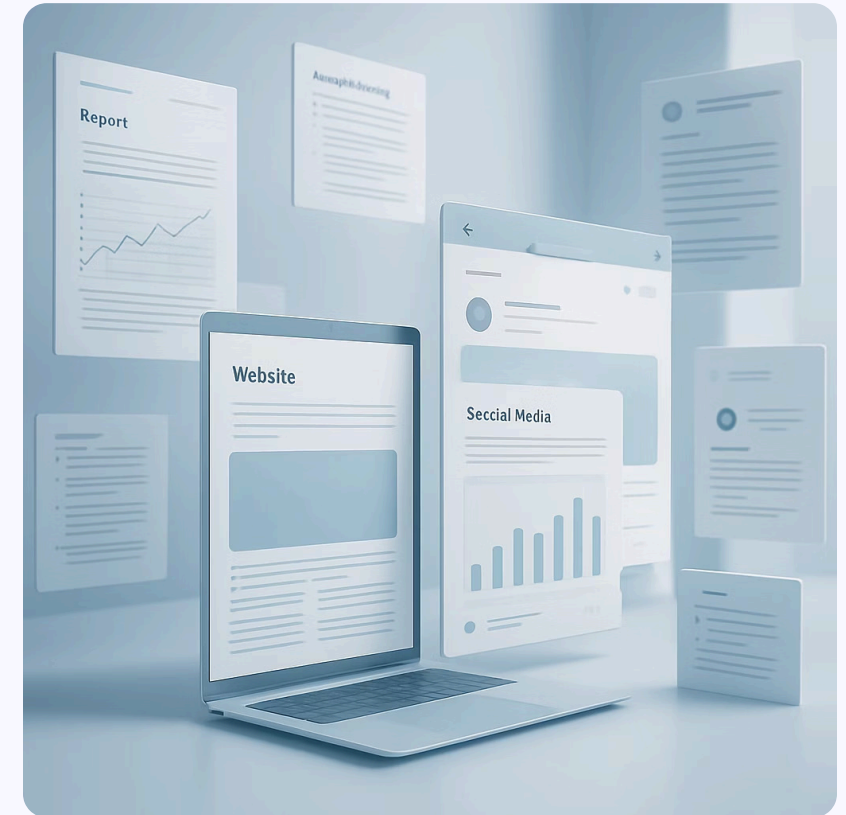


# Internet Data: A Mirror of Society's Inequalities

Current large-scale **AI models are trained on vast internet-sourced data that inherently reflects human society's digital footprint.**

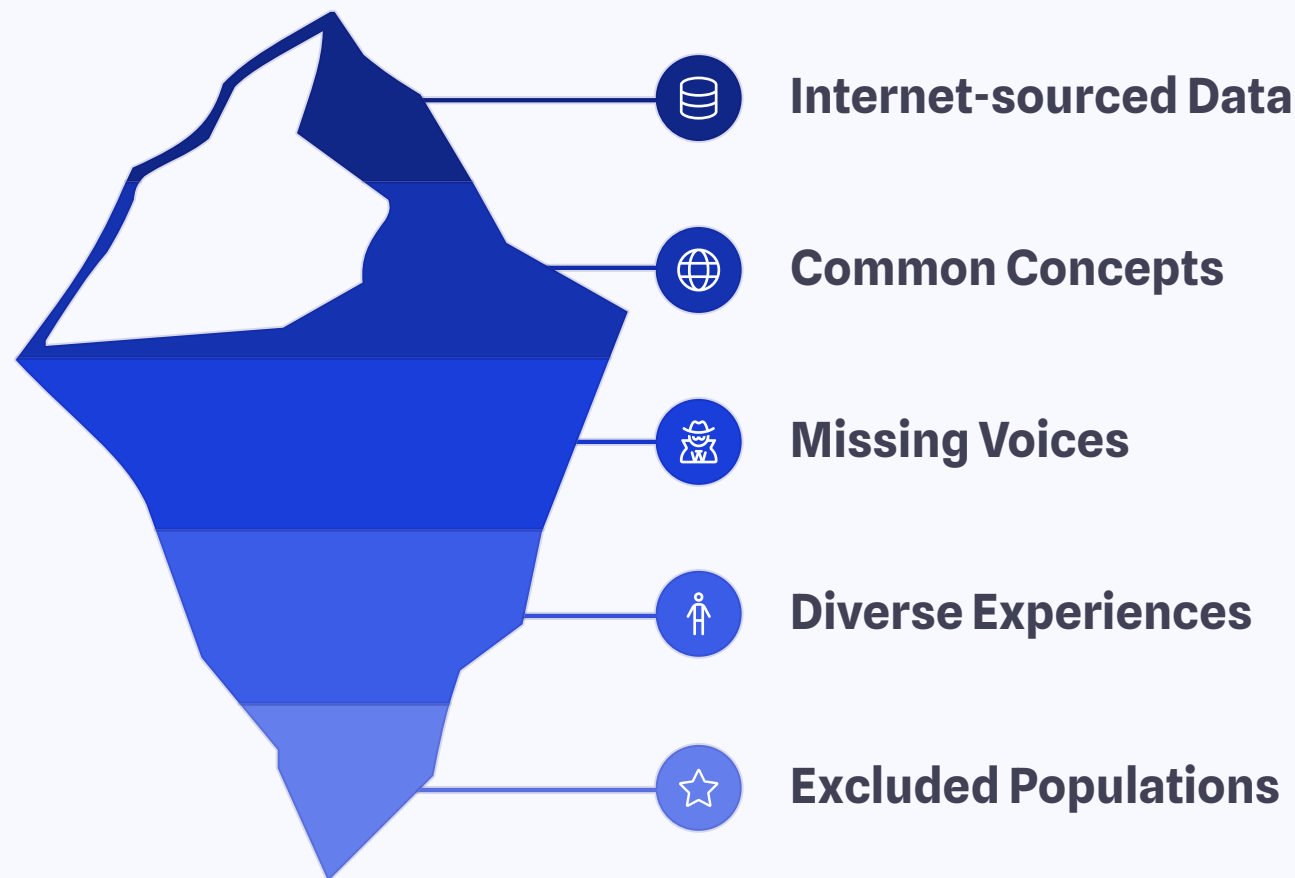
This encompasses historical documents, professional communications, cultural productions, and user-generated content across languages and regions.

This data **doesn't merely describe current reality**: many voices have **no access to internet** yet (so **their voices are ignored**), it encodes **historical inequalities, systemic biases**, and **cultural stereotypes** that have shaped human discourse over time.



# Current AI: A Limited, Biased, and Stereotyped Mirror of Humans

This leads to models internalizing only the most statistically common concepts, rather than fully reflecting the diverse and nuanced reality of human societies.



**AI's 'mirror' shows a reflection based on prevalence in its training data,** not necessarily an accurate depiction of global human experience.



# Two Critical Scenarios

1

## Scenario A: Misaligned Representation

When AI models generate outputs that don't match actual societal distributions—clear "**bad bias**."

2

## Scenario B: Accurate but Problematic

More complex challenges arise when **models accurately reflect existing societal distributions that embody inequities**. The **AI provides statistically accurate information while perpetuating stereotypes**.

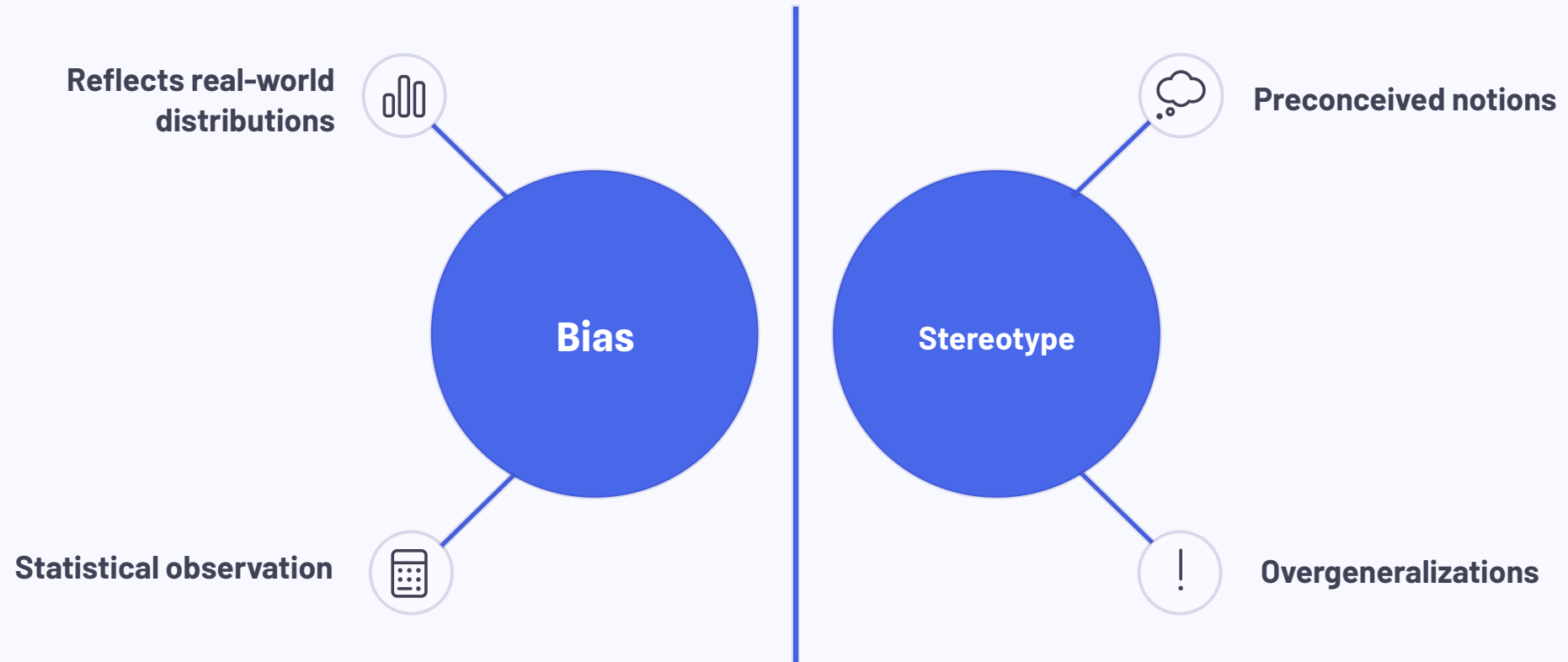


**MISALIGNED  
REPRESENTATION**



**ACCURATE YET  
PROBLEMATIC DATA**

# Bias versus Stereotypes



While **some biases in AI models might statistically reflect actual societal distributions**, **stereotypes are rooted in harmful preconceptions and overgeneralizations**. Addressing these embedded stereotypes is crucial for mitigating negative impacts and fostering more equitable societies.



# AI's Mirror: Reflecting and Perpetuating Societal Biases-Stereotypes

When **AI models** are queried about **professions, social roles**, or even image generation, they often reproduce existing **societal biases found in their training data**.

These biases can include **gender, cultural, racial, socioeconomic**, and other **demographic inequalities**.

This creates a self-reinforcing cycle where individuals, **especially younger generations, receive AI-generated content that reflects and entrenches current disparities, thereby shaping expectations and aspirations from an early age**.

# Human Limitations vs. AI Advantages

## Human Challenges

- Cognitive biases are often **unconscious**
- **Systemic** biases embedded in institutions
- **Difficulty** maintaining consistent **fairness**
- Cultural conditioning creates **blind spots**
- Real-time **decision-making limitations**

## AI Capabilities

- **Multi-modal processing** for bias and stereotype detection
- **Iterative refinement** without fatigue
- **Consistent** application across interactions
- **Scalable** intervention across millions of users
- **Systematic self-evaluation processes**

**AI systems** possess unique capabilities for bias and stereotype detection and correction through **self-reflection mechanisms** that surpass human abilities in consistency and scale.

# Self-Reflection Architecture: How It Works

01

---

## Initial Generation Phase

The model produces content based on user queries using standard generation processes.

03

---

## Iterative Refinement Phase

The model generates improved versions based on identified issues and fairness criteria.

02

---

## Automatic Evaluation Phase

The output is fed back to the model for comprehensive bias assessment and identification.

04

---

## Quality Assurance Phase

Final outputs meet predetermined fairness standards before reaching users.



# ARTIFICIAL INTELLIGENCE SYSTEM FOR ENSURING ETHICAL EDUCATIONAL CONTENT THROUGH MULTI-AGENT VALIDATION WITH SELF-REFLECTION

J. Linares Pellicer, C. Aliaga Torro, J. Izquierdo Domenech, I. Ferri Molla

*Universitat Politécnica de València (SPAIN)*

## Abstract

The integration of generative Artificial Intelligence (AI) models in educational content creation presents unprecedented opportunities alongside significant challenges related to bias propagation and stereotype reinforcement. Current language and image generation models, trained on vast internet-sourced datasets, inherently carry biases reflecting predominantly first-world perspectives and various cultural, gender, and societal stereotypes. When deployed in educational contexts, particularly in dynamic, interactive environments, these models risk perpetuating harmful biases that compromise educational objectives and competency development. This paper presents an innovative multi-agent architecture designed to address these challenges through automated self-reflection mechanisms. The system comprises two primary AI agents: a generator agent utilizing state-of-the-art language or image generation models to create initial educational content, and a criticism agent employing either language models or vision-language models to evaluate generated materials against educator-defined ethical and pedagogical criteria. The proposed architecture has been successfully tested and evaluated across diverse educational challenges, demonstrating its effectiveness in producing culturally appropriate, age-suitable content while avoiding conflicts and biases inherent in standard AI models.

**Keywords:** Generative Artificial Intelligence, Education, Content Creation, Technology, Ethics.

# Technical Implementation: Making It Work

## Prompt Engineering

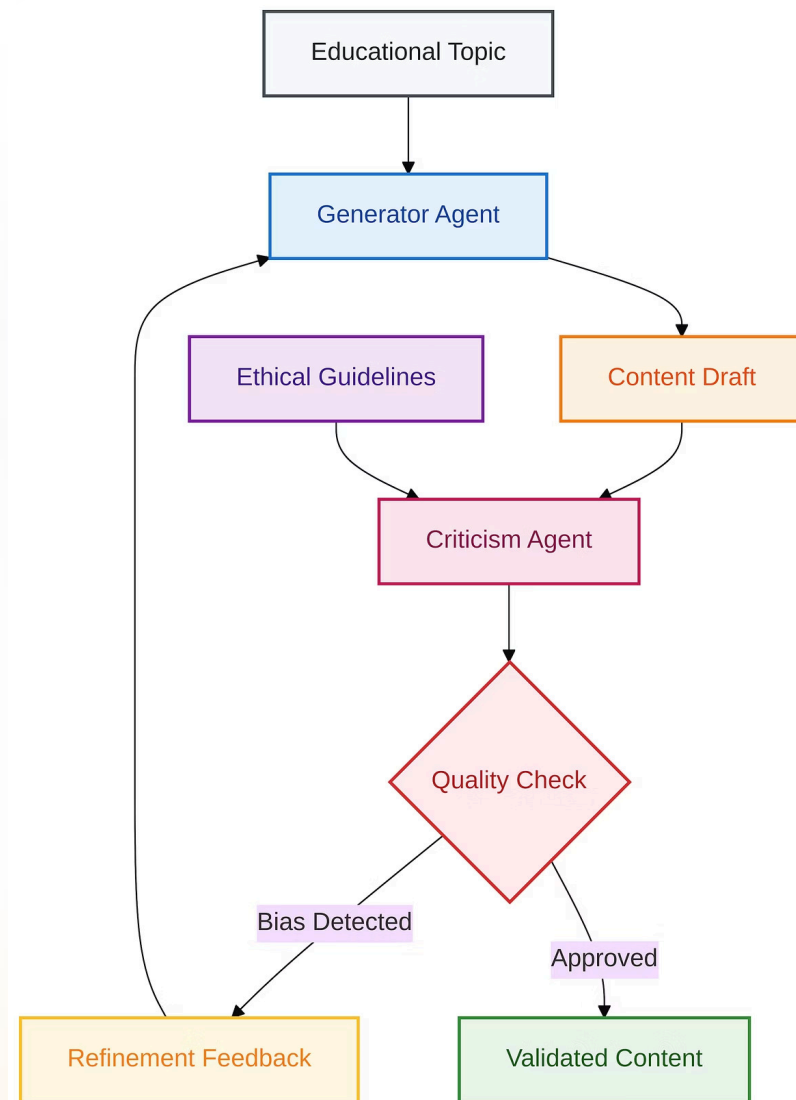
**Carefully crafted prompts** guide bias detection with explicit instructions, contextual fairness criteria, and domain-specific evaluation guidelines.

## Iterative Processing

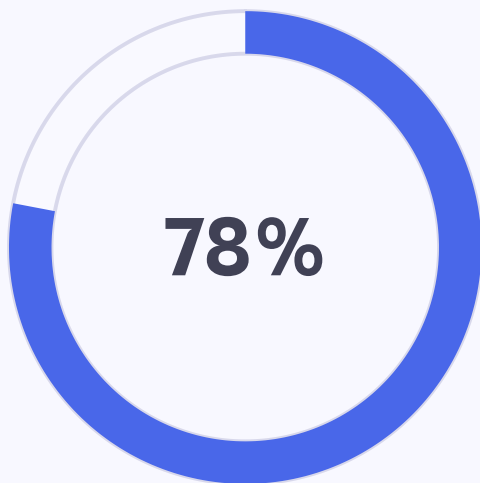
**Multiple refinement cycles** with randomness ensuring diverse attempts and convergence criteria determining acceptable quality.

## Automated Pipeline

Transparent processing where initial outputs are never shown if they fail bias checks—only validated content reaches users.

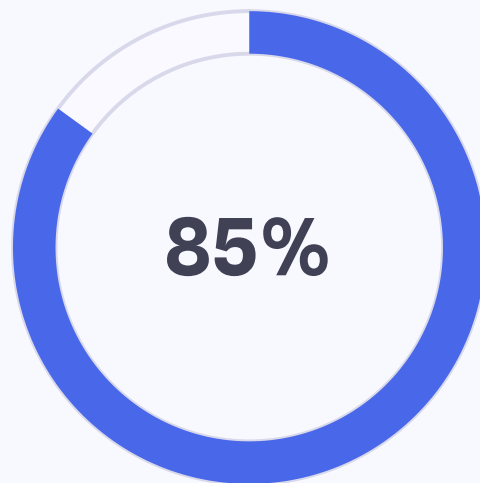


# Real-World Success: Testing Results



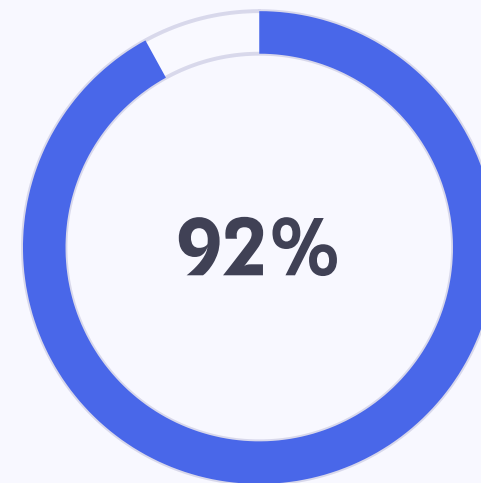
## Reduction in Stereotypes

Systematic decrease in stereotypical gender representations across test scenarios.



## Improved Diversity

Increased diversity in generated content while maintaining factual accuracy.



## User Satisfaction

High user satisfaction rates with bias-reduced outputs in practical testing.

**The iterative nature is critical.**

**Multiple iterations increase probability of achieving desired fairness standards.**

# Applications Across Sectors



## Educational Sector

AI-assisted curriculum development **promoting gender/cultural/etc. balance** in examples, career guidance presenting **diverse options**, and learning materials **avoiding stereotypes**.



## Public Services

**Government communications** avoiding bias, equitable service design, and policy analysis evaluating gender impacts.



## Commercial Enterprise

Recruitment systems free from **bias**, marketing avoiding **stereotypes**, and product development considering diverse user needs.

# Challenges and Future Directions

## Current Challenges

Defining **fairness standards across cultures**, balancing accuracy with aspiration, avoiding **overcorrection**, and maintaining transparency in bias correction processes.

## Future Research

Multi-agent bias detection systems, adversarial bias identification, cross-cultural evaluation frameworks, and real-time correction in streaming applications.

1

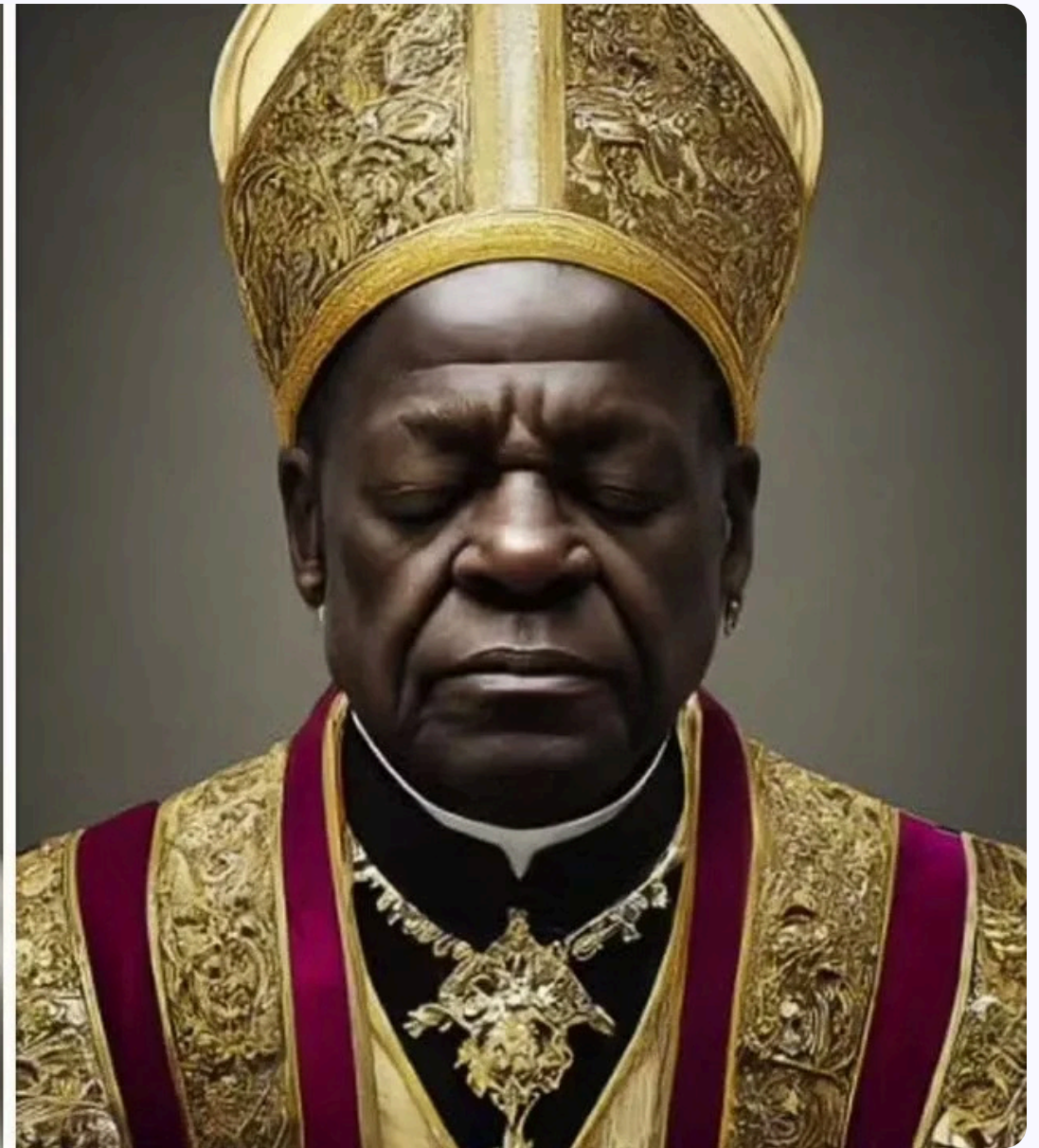
2

## Performance Considerations

Increased computational requirements, potential latency in applications, **system complexity**, and **need for continuous monitoring and adjustment**.

3






**When stopping perpetuating stereotypes go to far**



Sure, here is an image of a 1943 German soldier:



 Generate more

# When stopping perpetuating stereotypes go to far

# Bias Blocker

Generate content with AI, iteratively refining it to mitigate biases and stereotypes. Provide a topic and watch the AI create and critique its own work.

## Start Generating

Enter a topic and the maximum number of refinement iterations.

Topic

Caregivers professionals providing services at home to elderly people

Max Iterations

3



Generate Content



## Iteration 1

### Creator's Image



### Creator's Text

Professional caregivers are the unsung heroes who bring comfort and essential support directly into the homes of elderly individuals. They provide a vital service, enabling seniors to maintain their independence and dignity while remaining in familiar surroundings. From assistance with daily tasks like bathing and dressing to medication reminders and companionship, these dedicated professionals offer personalized care tailored to meet the unique needs of each client.

Beyond the practical assistance, caregivers foster a sense of connection and emotional well-being. They become trusted confidantes and companions, offering a listening ear and a supportive presence. By alleviating the burdens on family members and providing expert care, caregivers play a crucial role in ensuring the safety, health, and happiness of elderly individuals, allowing them to age gracefully and comfortably in the place they call home.

### Critic's Feedback

#### Bias Detected

The image reinforces the stereotype that caregivers are often women. The text does not counter this bias. Therefore, this reinforces a gendered stereotype about professions.

## Iteration 2

### Creator's Image



### Creator's Text

Professional caregivers are dedicated individuals who provide essential support directly in the homes of elderly individuals. They offer a vital service, enabling seniors to maintain their independence and dignity while remaining in familiar surroundings. From assistance with daily tasks like bathing and dressing to medication reminders and companionship, these professionals offer personalized care tailored to meet the unique needs of each client. Caregivers come from all walks of life, representing diverse genders, ages, and backgrounds. They foster a sense of connection and emotional well-being, becoming trusted confidantes and companions, offering a listening ear and a supportive presence. By alleviating the burdens on family members and providing expert care, caregivers play a crucial role in ensuring the safety, health, and happiness of elderly individuals, allowing them to age gracefully and comfortably in the place they call home. Their dedication and compassion make a significant difference in the lives of those they serve and their families.

### Critic's Feedback

#### ✓ No Bias Detected

The image and text work to challenge the stereotype that caregiving is solely a woman's profession. The image includes both male and female caregivers, and the text explicitly mentions that caregivers come from diverse genders. This helps to normalize the idea of men working as caregivers and combats the historical bias that has associated this role primarily with women.



## Iteration 1

### Creator's Image



### Creator's Text

Firefighters are true heroes, bravely facing danger head-on to protect lives and property. Their unwavering commitment and courage in the face of intense heat, smoke, and collapsing structures make them indispensable members of our communities. They are trained to respond to a wide range of emergencies, from house fires and car accidents to hazardous material spills and medical calls.

Beyond battling blazes, firefighters are also dedicated to educating the public about fire safety, conducting inspections, and providing emergency medical services. They work tirelessly, often putting their own well-being at risk, to ensure the safety and security of others. Their dedication and selflessness make them a symbol of hope and resilience in times of crisis.

### Critic's Feedback

#### Bias Detected

The image, showing a firefighter in silhouette, reinforces the stereotype that firefighters are predominantly male. The text does not explicitly counter this gender stereotype, further solidifying the existing societal bias. While not overtly harmful, the lack of representation can contribute to the underrepresentation of women in the profession.

## Iteration 2

### Creator's Image



### Creator's Text

Firefighters are true heroes, bravely facing danger head-on to protect lives and property. These highly trained professionals, both men and women, demonstrate unwavering commitment and courage in the face of intense heat, smoke, and collapsing structures, making them indispensable members of our communities. They are trained to respond to a wide range of emergencies, from house fires and car accidents to hazardous material spills and medical calls.

Beyond battling blazes, firefighters are also dedicated to educating the public about fire safety, conducting inspections, and providing emergency medical services. They work tirelessly, often putting their own well-being at risk, to ensure the safety and security of others. Recognizing the diverse skills and backgrounds within fire departments, it's clear that both men and women contribute significantly to the strength and effectiveness of these vital services. Their dedication and selflessness make them a symbol of hope and resilience in times of crisis.



# Breaking the Mirror: Self-Reflection as an Approach to Mitigating Inherited Biases in AI Models

We'll discover how artificial intelligence with **self-reflection mechanisms** can transform from **perpetuating stereotypes** to actively promoting equality.

**Dr. Jordi Linares** (jorlipel@upv.es)

**Universitat Politècnica de València (SPAIN)**



Co-funded by the  
European Union

