Féidearthachtaí as Cuimse
Infinite Possibilities

Understanding Gender Bias in AI

Exploring Causes, Consequences, and Mitigation Strategies

Dympna O'Sullivan

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

3rd International EUGAIN Summer Training School: Gender & Sustainability

Outline

- Introduction
- Bias and Gender Bias in AI
- Impact of Bias
- Real World Examples
- Mitigation Strategies
- Group Activity
- Wrap Up

About me

- Academic Lead
  - Digital Futures Research Hub, TU Dublin
- BSc and MSc in Computer Science
- Research in:
  - AI applied to Healthcare
  - AI Ethics

Relevant projects and resources

- Ethics4EU:
  - Ethics4EU – Resources to teach Computer Ethics in Computer Science and Engineering programmes (ascnet.ie)

- Inclusion4EU
  - Inclusion4EU – Co-Design for Inclusion in Software Development Design (ascnet.ie)

AI Bias in the News

PC Tablet · 8d

Google's AI Image Generator Addresses Bias: Tackling the White People Glitch

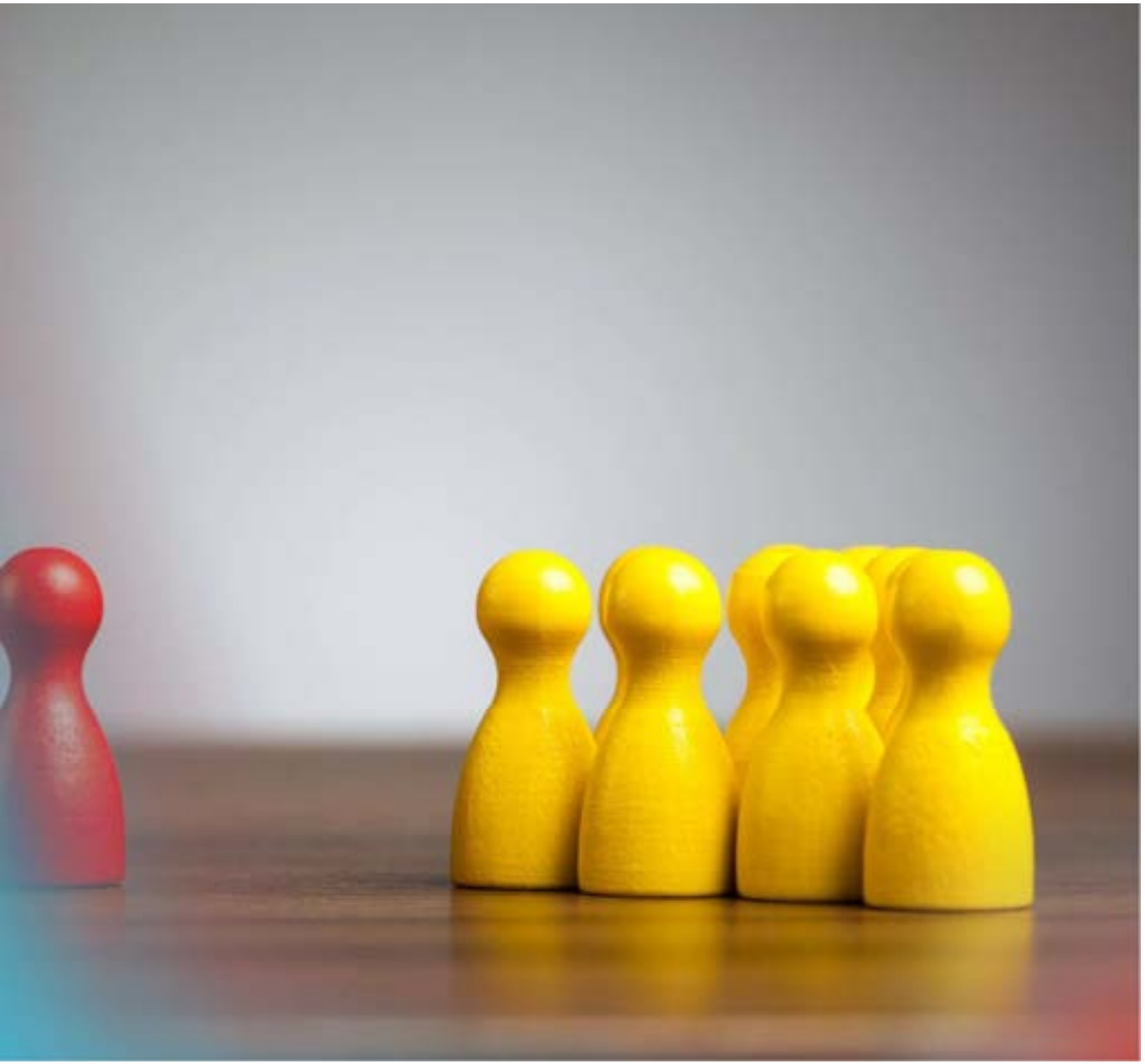
Google addresses bias in its image-generating AI, Imagen, after it failed to depict white people accurately. The company is ...

The Telegraph · 1d

Amazon accused of bias after Alexa gives users reasons to vote for Harris and not Trump

Users on social media accuse company of election interference as its virtual assistant appears to favour Democratic candidate ...

TU DUBLIN



Bias

- Inclination or prejudice for or against one person or group, especially in a way considered to be unfair

How does bias seep into technology?

- Orlikowski posits that technologies are:
  - ***"products of their time and organizational context"*** which ***"will reflect the knowledge, materials, interests, and conditions at a given locus in history"***.
  - As technology is ***"both structurally and socially constructed",*** it both mirrors the implicit biases of its creators, while also gaining new meanings and functions—and potentially biases—through repeated and widespread use.
  - When governments employ these technologies, from search engines to recruitment software, they may be ***unwittingly amplifying such biases, which in turn may influence outcomes from policy to hiring decisions.***

Gender Bias

- Gender bias, according to the European Institute for Gender Equality (2023), refers to “prejudiced actions or thoughts based on the gender-based perception that women are not equal to men in rights and dignity”.

Gender Bias in AI

- Gender bias in AI refers to the systematic favouring of one gender over another by AI systems, leading to unfair outcomes.

- This bias can manifest in various ways, from skewed data sets to biased algorithmic decision-making.

Importance of Addressing Bias

- AI systems are increasingly used in critical areas like hiring, healthcare, and law enforcement
- Addressing gender bias is crucial to ensure fairness, equality, and ethical AI deployment
- Addressing bias is essential for building public trust in AI

Causes of Gender Bias in AI

- Data
- Models
- Algorithm Design

Bias in Data Collection

- **Historical Biases:** AI systems trained on data that reflect societal biases can learn and perpetuate those biases
- **Skewed Data Sets:** If data sets underrepresent certain genders, the AI model may not perform well for those groups

Gender Bias Data Example

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi et al., 2016)

Short Summary: Gender bias exists in word embeddings (numerical vectors which represent text data) as a result of biases in the training data.

Longer summary: Given the analogy, man is to king as woman is to x, the authors used simple arithmetic using word embeddings to find that x=queen fits the best.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Subtracting the vector representations for "man" from "woman" results in a similar value as subtracting the vector representations for "king" and "queen". From Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

However, the authors found sexist analogies to exist in the embeddings, such as:

- He is to carpentry as she is to sewing
- Father is to doctor as mother is to nurse
- Man is to computer programmer as woman is to homemaker

Bias in Model Trainings

- **Training Algorithms:** If models are trained on biased data, they learn to reproduce these biases in their predictions and decisions.

- **Feature Selection:** The features chosen to train the model might inherently favor one gender, leading to biased outcomes.

Bias in Model Training example

"Compassionate manager" by Stable Diffusion.

"Manager" by Stable Diffusion.

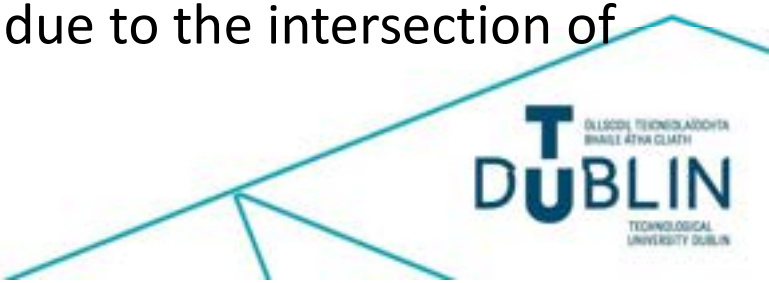
Bias in Algorithm Design

- **Objective Functions:** Algorithms designed without considering fairness can optimize for accuracy at the expense of equity.
- **Lack of Diversity in Design Teams:** Homogeneous teams may unintentionally embed their own biases into AI systems.

Intersectionality

- Intersectionality is a framework that examines how various aspects of a person's identity (such as race, gender, class, sexuality, and ability) intersect and interact, creating unique experiences of privilege or oppression.

  - **Compounded Discrimination:** An AI system might disproportionately disadvantage individuals who belong to multiple marginalized groups.
    - For example, a Black woman might face more severe bias in an AI system than a White woman or a Black man due to the intersection of racial and gender biases.



Social and Ethical Implications of Gender Bias in AI

- Reinforcement of Stereotypes
  - AI systems can amplify and reinforce harmful gender stereotypes, further entrenching societal biases.
- Exclusion/marginalization
  - Biased AI can lead to the exclusion of certain groups from opportunities, such as job offers or financial services.
- Unequal Opportunities
  - Gender bias in AI can perpetuate or amplify inequality by disproportionately benefiting one gender over others in critical areas like healthcare, education, and employment.


Real World Examples

Healthcare

- Problems arising from male-dominated training data
  - Inequitable Treatment:
    - Medical treatments that are less effective or even harmful for certain genders, e.g. medications or therapies tested primarily on male subjects may not work as well or have adverse effects in women.
  - Diagnostic Errors:
    - Misdiagnosis or delayed diagnosis for conditions that manifest differently across genders.

Cardiovascular AI Tools

- A 2019 study published in the *Journal of the American Heart Association* highlighted that AI models used for predicting heart attack risk were less accurate for women.
- Women are more likely to experience atypical symptoms of heart disease, such as shortness of breath, nausea, or back pain, rather than the classic chest pain more common in men.
- AI systems trained on male data may not recognize these symptoms as indicators of heart disease, leading to missed or delayed diagnoses.

Finance

- Credit scoring systems offering lower credit limits to women despite similar financial profiles as men
  - Historical biases, such as women being offered lower credit or being less likely to receive loans
  - Feature Weighting, giving undue weight to factors that correlate with gender rather than financial behavior, such as employment type or marital status, which can disadvantage women

The Apple Card Incident

- Numerous reports surfaced where women, including prominent individuals like tech entrepreneur David Heinemeier Hansson’s wife, were offered significantly lower credit limits than their male counterparts.

- In one case, Hansson received a credit limit 20 times higher than his wife, even though they shared financial accounts and had similar credit histories.

Recruitment

- Gender bias in recruitment refers to the unfair treatment or discrimination against individuals based on their gender during the hiring process.

- This bias can manifest at various stages, including job postings, resume screening, interviews, and final hiring decisions.

Amazon

- In the mid-2010s, Amazon developed an AI-based system to help automate the process of reviewing job applicants' resumes (CVs).
- The goal was to streamline hiring by using machine learning to sift through large volumes of resumes and identify the best candidates.
- However, the system quickly became notorious for exhibiting gender bias, particularly against women.

Mitigation Strategies

- Diverse Data Sets
- Inclusive Design Practices
- Fairness Algorithms

Diverse Datasets

- **Inclusive Data Collection:** Ensure that data sets used to train AI models include diverse and representative samples from all genders.

- **Bias Audits:** Regularly audit data sets for potential biases and address any imbalances
  - Data Sources
  - Data Quality
  - Data Labelling
  - Assess Feature Relevance
  - Fairness metrics

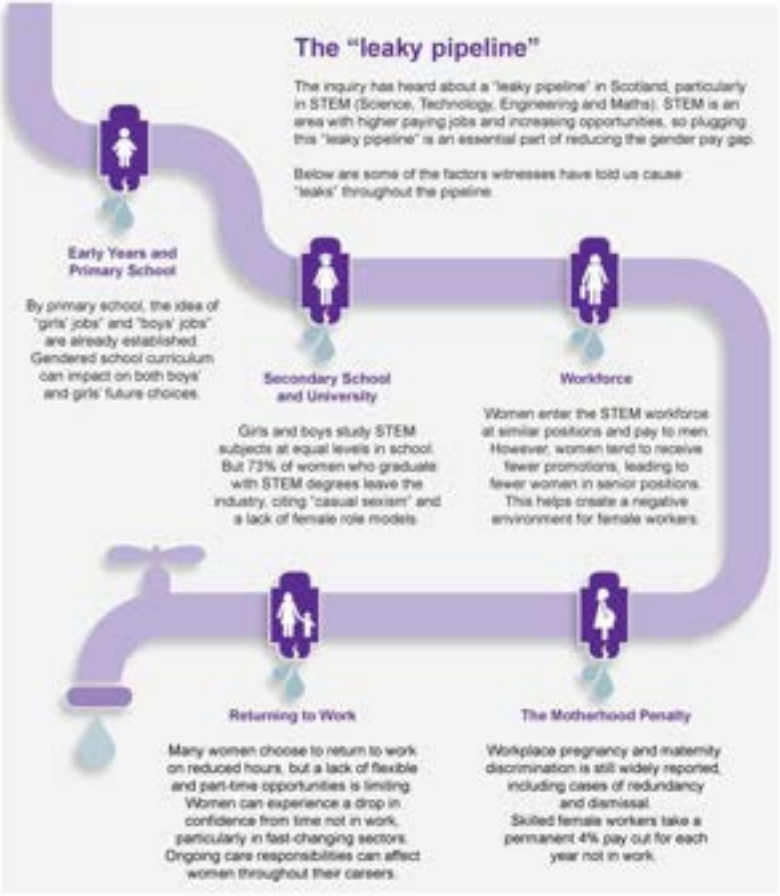
Fairness Algorithms

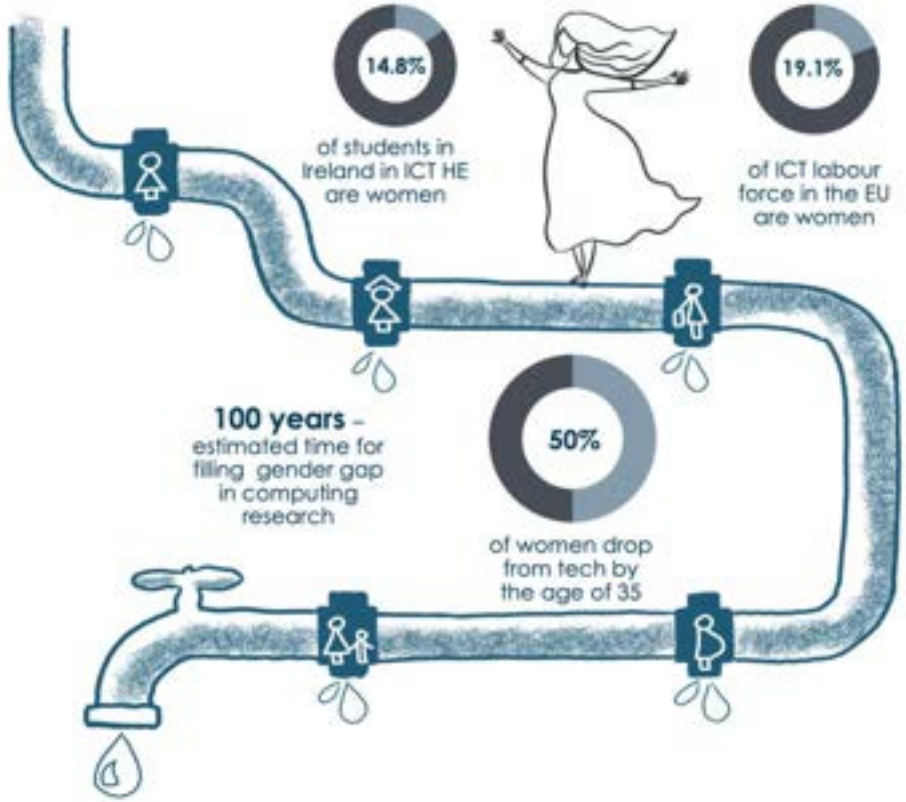
- Pre-processing algorithms
  - Re-weighting
  - Re-sampling
- In-processing algorithms
  - Adversarial debiasing
- Post-processing algorithms
  - Calibrated Equalized Odds

Inclusive Design Practices

- **Fairness by Design:** Incorporate fairness considerations into the design phase of AI systems to proactively address potential biases.

- **Diverse Teams:** Foster diversity in AI development teams to bring different perspectives and reduce the risk of embedding bias.





United Nations: AI Gender Bias is No Glitch in the System

Legislation

- Discrimination based on gender is illegal in many jurisdictions
  - But biases can unintentionally be embedded in technology and AI systems, leading to outcomes that may effectively discriminate against certain groups.
  - This highlights the paradox where legal protections exist, yet technological systems may still perpetuate inequality due to inherent biases in their design, data, or implementation.

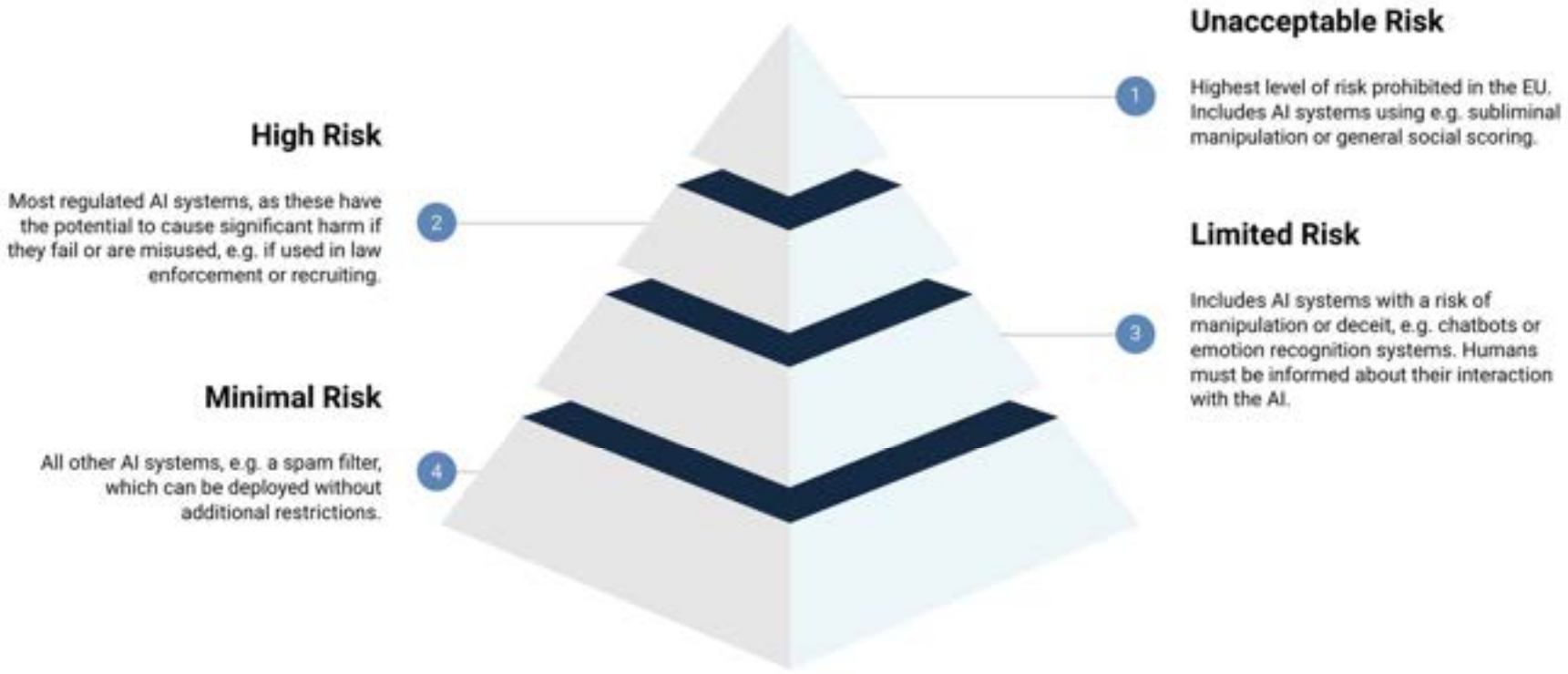
The EU AI Act

- A legal framework governing the sale and use of artificial intelligence in the EU
  - Official purpose is to ensure the proper functioning of the EU single market by setting consistent standards for AI systems across EU member states

Core Principles of the Act

- Human-centric AI:
  - How AI systems should be developed to benefit individuals and society at large, with human interests at the forefront
- Transparency:
  - Emphasizes the importance of clear, understandable, and explainable AI decision-making processes to build trust
- Accountability:
  - Discusses the responsibility of developers and users in ensuring the responsible use of AI systems and mitigating potential harm

Risk-Based Classification of AI Systems



Compliance and Penalties

- Member states will set up conformity assessment bodies in assessing and certifying AI systems' compliance with the EU AI Act

- The heftiest fines are imposed for violating the prohibition of specific AI systems, up to 40M EUR or 7% turnover

- The lowest penalties are for providing incorrect, incomplete or misleading information, up to 5M EUR or 1% of annual worldwide turnover

Bias and Sustainability

- Gender bias in AI is related to sustainability because both involve
  - creating systems and societies that are fair, inclusive, and equitable for all

Bias and Sustainability

- Gender bias in AI:
  - Undermines social sustainability by perpetuating inequalities and discrimination
  - Can limit economic opportunities for women and other marginalized genders, reinforcing economic disparities

- Sustainability in technology includes the ethical development of AI systems that do not harm any group.

Conclusion

- Gender bias in AI is a significant challenge with real-world consequences
- Bias can be introduced at multiple stages, from data collection to algorithm design
- Mitigating bias requires diverse data, inclusive design practices, and specialized algorithms

Group Activity

- AI-based healthcare system
- Roles
  - **Healthcare Provider (Doctor/Nurse)**
  - **Patient (Female)**
  - **AI Developer/Engineer**
  - **Hospital Administrator**
  - **Policy Maker/Regulator**

- AI-based recruitment system
- Roles
  - **HR Manager/Recruiter**
  - **Job Applicant (Female)**
  - **AI Developer/Engineer**
  - **Chief Diversity Officer**
  - **Company CEO/Executive**
  - **Employment Lawyer**