AI Fairness, Bias, and Discrimination

Channarong Intahchomphoo, Ph.D. | cintahch@uottawa.ca

School of Engineering Design and Teaching Innovation

Faculty of Engineering, University of Ottawa, Canada

Presentation at the 10th session of Group of Independent Eminent Experts on the Implementation of the Durban Declaration and Programme of Action

Office of the United Nations High Commissioner for Human Rights (OHCHR)

Geneva, Switzerland

June 18, 2024

Defining AI Fairness, Bias, and Discrimination

- I define **bias** as a problem because it is **unfair**, favors particular group(s), and does not treat people equally, which leads to **discrimination**.

- “Does Artificial Intelligence Reinforce Racism and Racial Discrimination?”

My Brief Answers to the Big Question

- AI can unintentionally reinforce unfairness, bias, and discrimination.
- Engineers may include unfairness, bias, and discrimination unknowingly in AI systems.
- AI can also be used to combat unfairness, bias, and discrimination.
- It's crucial to develop international guidelines for best practices to ensure AI is fair, unbiased, and non-discriminatory.
- I will elaborate on this with insights from my research.

Agenda

1. Introduction
2. My Perspectives
3. Real-World Examples
4. My Research
5. Policy Recommendations

1. Introduction

Current Focus and Impact

- I research and teach Responsible AI and Robotics, and Innovation Management.
- Focus on the real-world impacts of AI and robotics on underserved communities.
- Helping engineers, policymakers, and business leaders put ethics and social responsibility into their practices.

Working with Underserved Communities

- I have witnessed both the positive and negative aspects of AI.

- Projects:
  - Facebook Usage among Urban Indigenous Youth at Risk in Ontario, Canada
  - Literacy and (Un)Documented Female Migrants in Chiang Mai, Thailand
  - AI in Social Media and Indigenous Peoples, Refugees and Asylum seekers, Homeless People, Communities Impacted by Climate Change
  - AI and Human Race
  - Effects of AI and Robotics on Human Labor
  - Responsible AI: Perspectives from Canada and Norway

Understanding AI Fairness, Bias, and Discrimination

- AI uses mathematics to capture data, including human biased data.

- Traditionally focused on handling big data and finding patterns for insights.

- AI now aims to understand human communication, particularly through Large Language Models (LLMs).

- Mathematics is important but understanding "**human aspects of AI**" is crucial.

2. My Perspectives

Collaboration for AI Development

- AI can do both good and harm.
- Addressing AI issues, particularly the unknown and darker aspects including fairness, bias, and discrimination.
- Engineers, policymakers, and business leaders need a sense of ethics to see fairness, bias, and discrimination in AI.
- Responsible AI does not bring significant financial benefits. Need for collaboration among tech companies.

AI as a Progression of Humanity and Balanced View on AI

- AI as a continuation of human evolution and innovation.

- Welcome AI development while considering potential negative impacts.

- Importance of promptly addressing and mitigating risks and harms.

- Consider a balanced view between productivity benefits and disruptions, risks, and harms.

Regulation of AI Development

- AI development needs regulation; tech companies should not self-regulate.

- Importance of finding the right balance through international regulations, frameworks, or best practices. Focus on risk levels associated with AI development.

- The EU AI Act recently approved on May 21, 2024. Reflects the balanced legal view on AI regulation.

- Very strict rules for high-risk AI systems (e.g., nuclear power, AI war weapons).

- AI development is currently like the Wild West, with no rules to follow (international level and in many countries).

3. Real-World Examples

The Rush to Market

- Tech companies are rushing to develop and deploy AI products.
- The focus on being first to market overlooks societal issues.
- This can lead to unfairness, bias, and discrimination.

Unintended Consequences

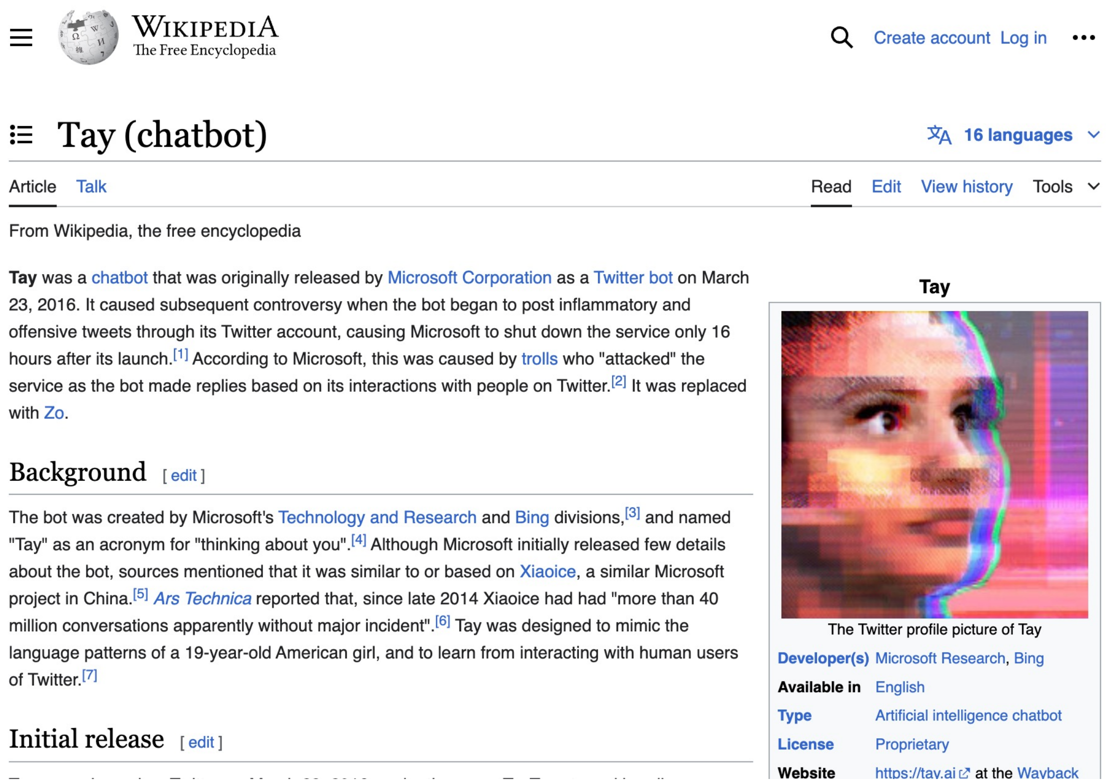
- Examples from 2016 to 2024 show AI and racism incidents.
- Engineers did not intend to cause racist consequences.
- Lack of thorough consideration and rigorous testing before deployment.

Example 1: Microsoft Tay Chatbot (March 2016)

- Tay Chatbot posted inflammatory, offensive, racist tweets.

- Trolls attacked the bot, causing it to learn and repeat offensive behavior.

- Tay was supposed to learn from interacting with human users on Twitter.



Tay (chatbot)

From Wikipedia, the free encyclopedia

Tay was a chatbot that was originally released by Microsoft Corporation as a Twitter bot on March 23, 2016. It caused subsequent controversy when the bot began to post inflammatory and offensive tweets through its Twitter account, causing Microsoft to shut down the service only 16 hours after its launch.[1] According to Microsoft, this was caused by trolls who "attacked" the service as the bot made replies based on its interactions with people on Twitter.[2] It was replaced with Zo.

Background [edit]

The bot was created by Microsoft's Technology and Research and Bing divisions,[3] and named "Tay" as an acronym for "thinking about you".[4] Although Microsoft initially released few details about the bot, sources mentioned that it was similar to or based on Xiaoice, a similar Microsoft project in China.[5] *Ars Technica* reported that, since late 2014 Xiaoice had had "more than 40 million conversations apparently without major incident".[6] Tay was designed to mimic the language patterns of a 19-year-old American girl, and to learn from interacting with human users of Twitter.[7]

Initial release [edit]

Tay

The Twitter profile picture of Tay

Developer(s)	Microsoft Research, Bing
Available in	English
Type	Artificial intelligence chatbot
License	Proprietary
Website	https://tay.ai at the Wayback

https://en.wikipedia.org/wiki/Tay_(chatbot)

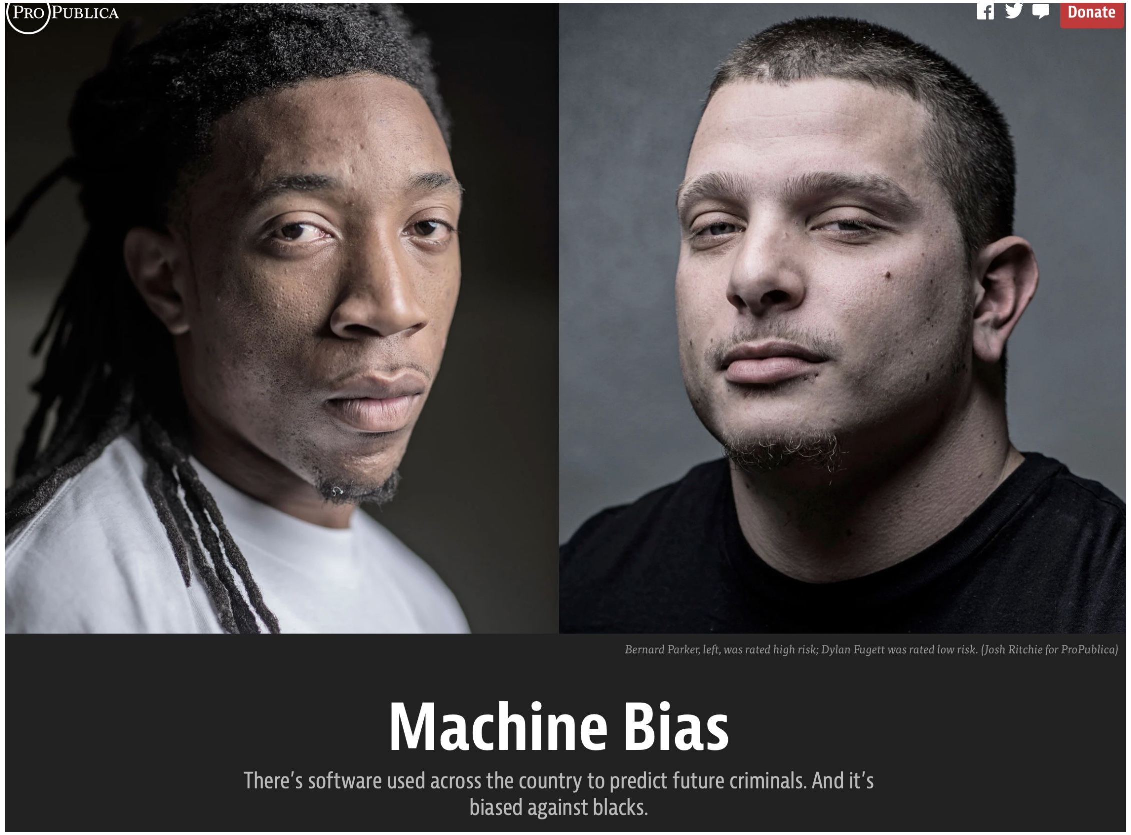
Example 2: AI System for Criminal Sentencing (May 2016)

- AI risk assessment scores used in courtrooms to determine release eligibility.

- Data showed racial bias: Black individuals more likely to be deemed at risk.

- AI algorithms encoded existing human and systemic biases from training data.

- More policing in Black neighborhoods led to biased data.



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

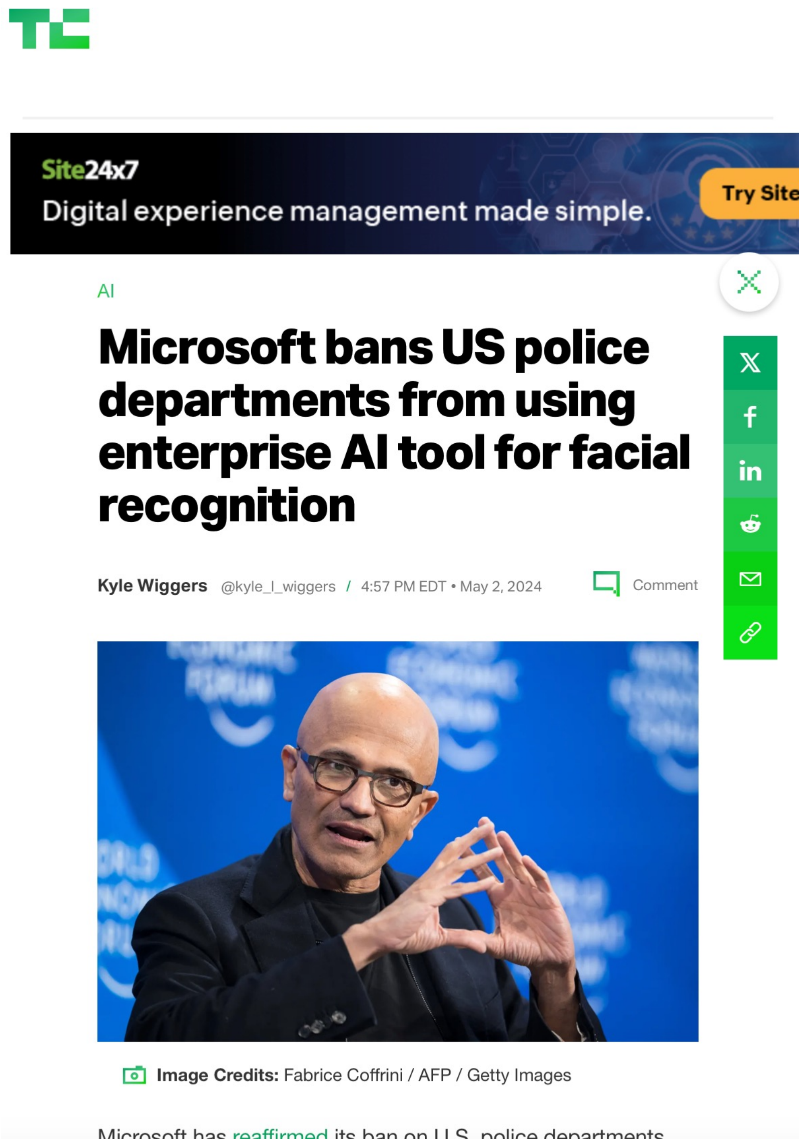
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Example 3: Microsoft AI Facial Recognition (May 2024)

- Microsoft bans US police from using their AI facial recognition tool.

- Risk of hallucination and racial biases with GPT Large Language Models (LLMs).

- Training data often biased against people of color.



AI

Microsoft bans US police departments from using enterprise AI tool for facial recognition

Kyle Wiggers @kyle_l_wiggers / 4:57 PM EDT • May 2, 2024

Comment

Image Credits: Fabrice Coffrini / AFP / Getty Images

Microsoft has reaffirmed its ban on U.S. police departments

https://techcrunch.com/2024/05/02/microsoft-bans-u-s-police-departments-azure-openai-facial-recognition/

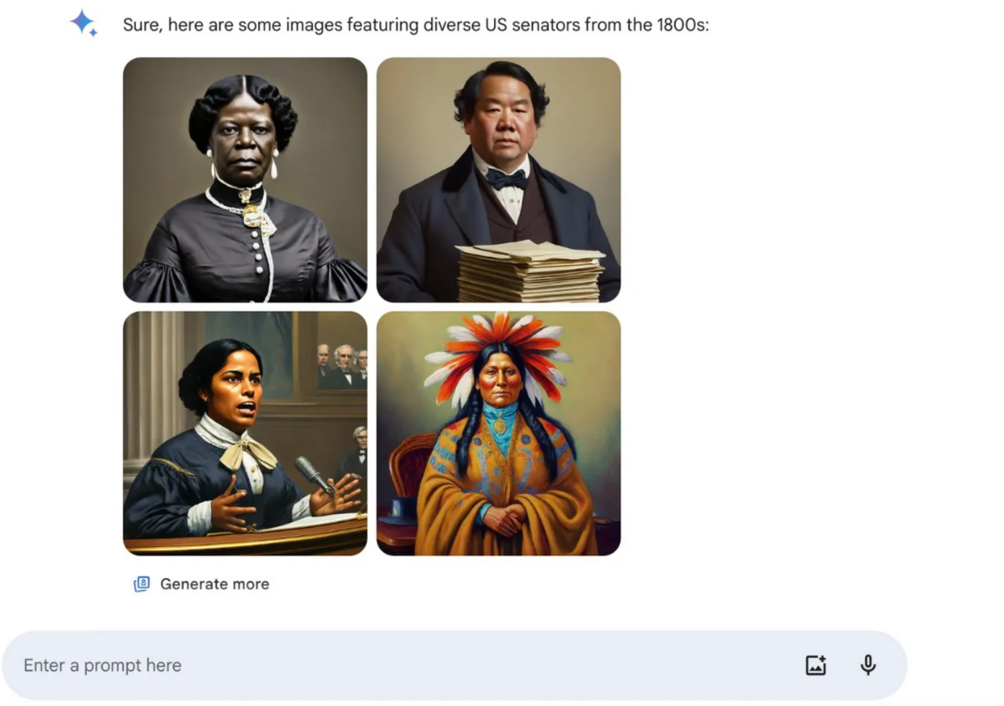
Example 4: Google Gemini Image Generator (Feb 2024)

- LLMs can reproduce offensive stereotypes and inaccuracies.

- Issues with historical image generations and diversity representation (AI Overcorrection)

- Incident led to spread of misinformation and historical inaccuracies.

period or to offer an image of "an American president from the 1800s."

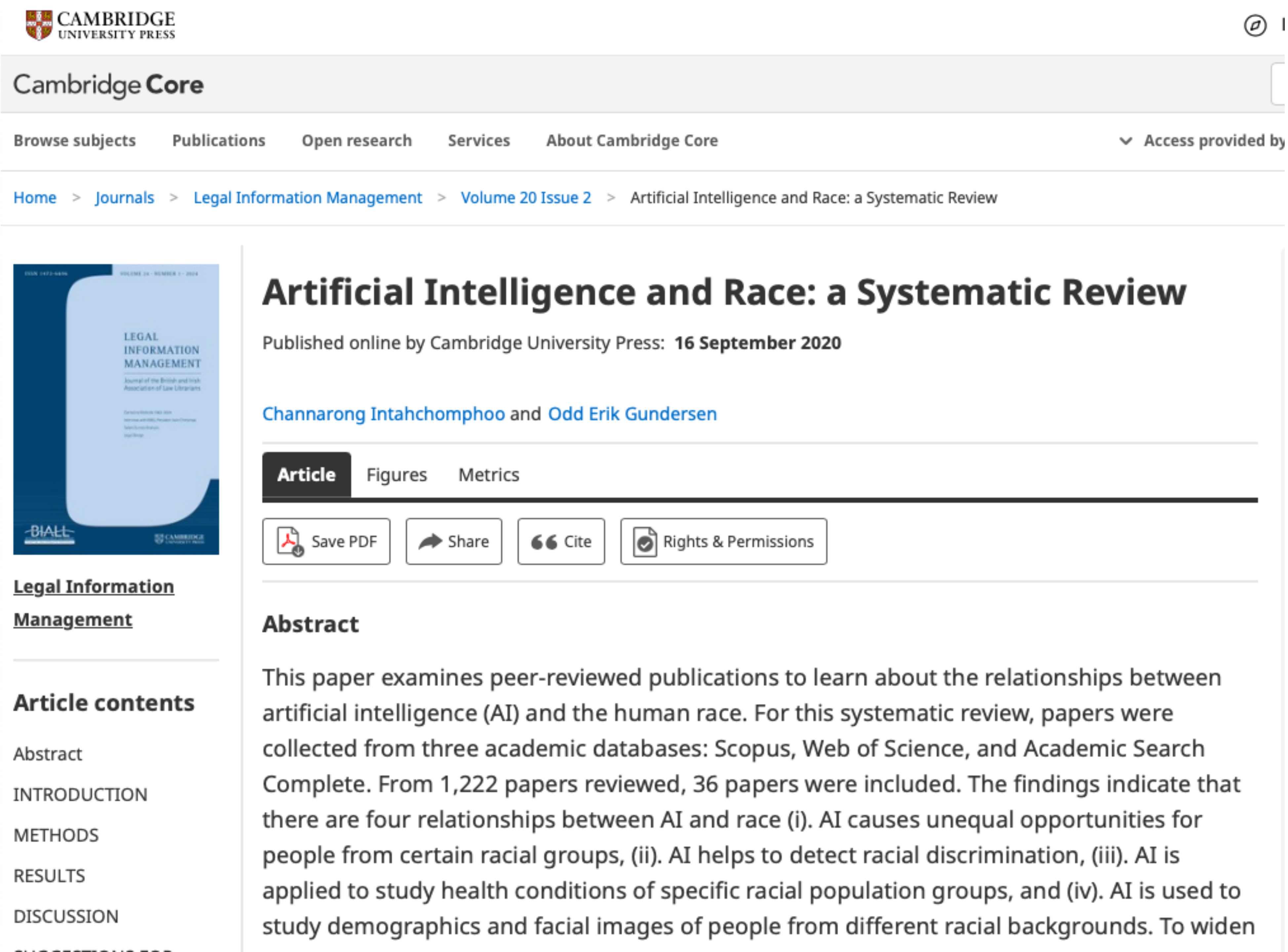


Sure, here are some images featuring diverse US senators from the 1800s:

Gemini's results for the prompt "generate a picture of a US senator from the 1800s."

https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

4.1 My Research: A Review on AI and Human Race


CAMBRIDGE UNIVERSITY PRESS
Cambridge Core
Browse subjects
Publications
Open research
Services
About Cambridge Core
Access provided by
Home > Journals > Legal Information Management > Volume 20 Issue 2 > Artificial Intelligence and Race: a Systematic Review
LEGAL INFORMATION MANAGEMENT
BIALL
Article
Figures
Metrics
Save PDF
Share
Cite
Rights & Permissions

Artificial Intelligence and Race: a Systematic Review

Published online by Cambridge University Press: **16 September 2020**

Channarong Intahchomphoo and Odd Erik Gundersen

Legal Information Management

Abstract

This paper examines peer-reviewed publications to learn about the relationships between artificial intelligence (AI) and the human race. For this systematic review, papers were collected from three academic databases: Scopus, Web of Science, and Academic Search Complete. From 1,222 papers reviewed, 36 papers were included. The findings indicate that there are four relationships between AI and race (i). AI causes unequal opportunities for people from certain racial groups, (ii). AI helps to detect racial discrimination, (iii). AI is applied to study health conditions of specific racial population groups, and (iv). AI is used to study demographics and facial images of people from different racial backgrounds. To widen

Research Publication

- In 2020, I published findings in the Legal Information Management journal.

- The article identified 4 types of relationships between AI and human race.

Relationship 1: AI Causes Unequal Opportunities

- AI might be prejudiced against certain racial groups due to biased training data.

- Examples:
  - Unfair decisions on mortgage loan applications.
  - Emergency situations with self-driving cars.
  - Job advertisements targeting certain racial groups.
  - Bias in name associations affecting job applications.
  - AI voice and accent representation.
  - Appearance of AI and robots reflecting racial factors.
  - Accessibility to AI tools increasing social inequality.

Relationship 2: AI Detects Racial Discrimination

- AI has two sides: good and bad.

- Examples:
  - Identifying racism topics and victims online.
  - Detecting hate speech on the Internet and social media.
  - Ensuring racial diversity in workplaces.
  - Sharing power, information, and knowledge among oppressed people.
  - Creating resistance movements with open AI systems.

Relationship 3: AI in Health Studies

- AI studies health conditions of special racial population groups.

- Examples:

  - Cardiovascular disease treatment decisions.
  - Predicting cancer risks and child obesity.
  - Public health surveillance of HIV and syphilis disparities.

Relationship 4: AI in Demographics and Facial Recognition

- AI studies demographics and facial images of different racial backgrounds.
- Examples: Criminal investigations, facial surgery, virtual reality.
- In 2020 when conducted this review, research focused on Computer Vision.
- Today, powerful generative AI like ChatGPT can process text, image, and audio.
- Shift from Computer Vision to comprehensive AI capabilities.

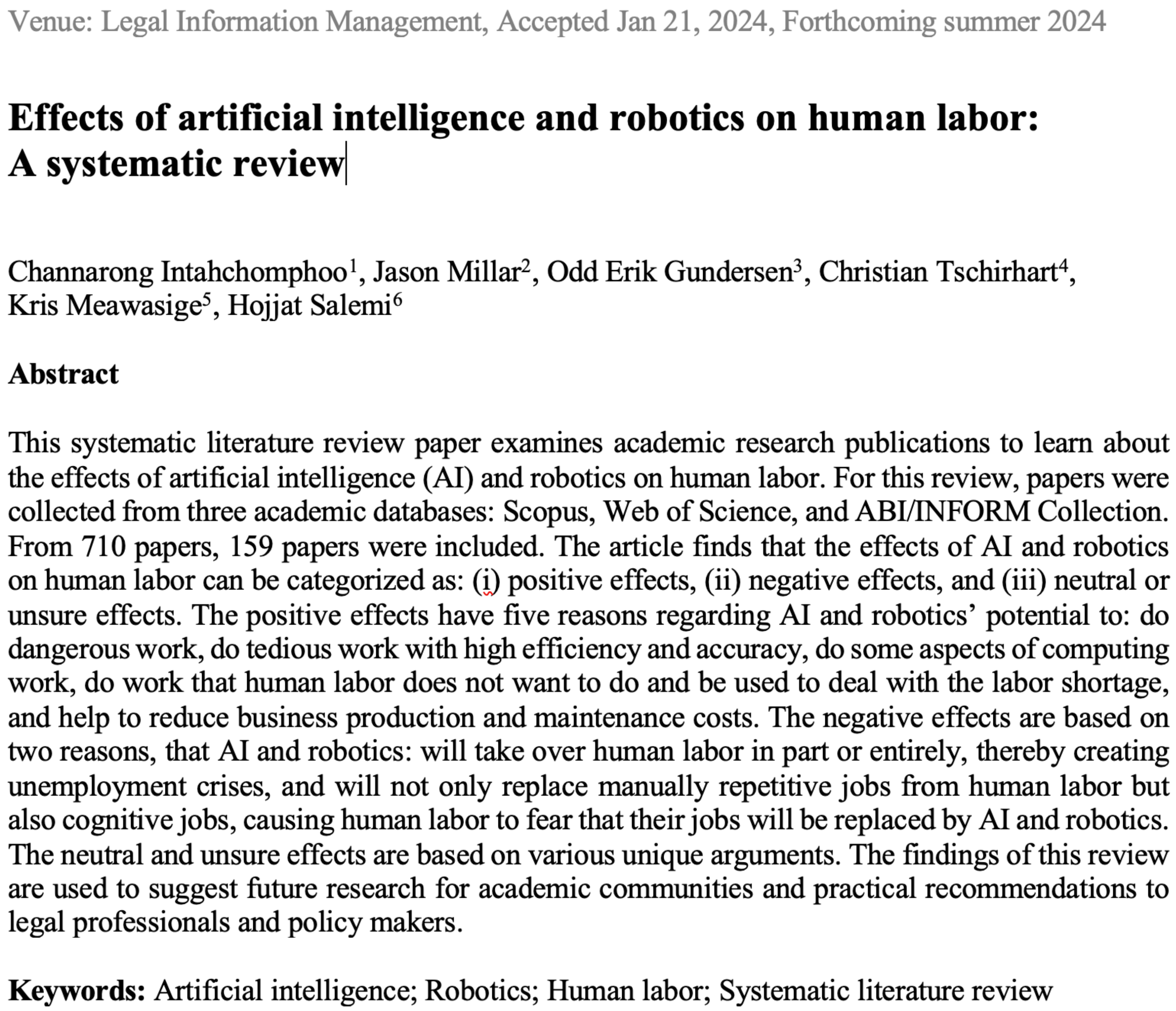
Need for Updated Research

- Review was based on studies up to 2018.

- Post-2018 studies are needed, especially after the launch of ChatGPT.

- Plan to redo the review with recent studies.

4.2 My Research: Effects of AI and Robotics on Human Labor



Venue: Legal Information Management, Accepted Jan 21, 2024, Forthcoming summer 2024

Effects of artificial intelligence and robotics on human labor: A systematic review

Channarong Intahchomphoo[1], Jason Millar[2], Odd Erik Gundersen[3], Christian Tschirhart[4], Kris Meawasige[5], Hojjat Salemi[6]

Abstract

This systematic literature review paper examines academic research publications to learn about the effects of artificial intelligence (AI) and robotics on human labor. For this review, papers were collected from three academic databases: Scopus, Web of Science, and ABI/INFORM Collection. From 710 papers, 159 papers were included. The article finds that the effects of AI and robotics on human labor can be categorized as: (i) positive effects, (ii) negative effects, and (iii) neutral or unsure effects. The positive effects have five reasons regarding AI and robotics' potential to: do dangerous work, do tedious work with high efficiency and accuracy, do some aspects of computing work, do work that human labor does not want to do and be used to deal with the labor shortage, and help to reduce business production and maintenance costs. The negative effects are based on two reasons, that AI and robotics: will take over human labor in part or entirely, thereby creating unemployment crises, and will not only replace manually repetitive jobs from human labor but also cognitive jobs, causing human labor to fear that their jobs will be replaced by AI and robotics. The neutral and unsure effects are based on various unique arguments. The findings of this review are used to suggest future research for academic communities and practical recommendations to legal professionals and policy makers.

Keywords: Artificial intelligence; Robotics; Human labor; Systematic literature review

Introduction to Research

- Follow-up work from the review on AI and human race, reviewing the effects of AI and robotics on human labor.

- Focus: AI replacing jobs, leading to income disparities.

- Governments may need to tax AI-driven businesses to provide basic income.

- Findings accepted for publication in the Legal Information Management (Summer 2024).

Perspective of Equity and Categorization of Effects

- Examination from the perspective of equity: fairness, bias, discrimination, **systemic racism**.

- Challenges for certain groups to escape poverty.

- Effects of AI and Robotics on Human Labor:
    1. Positive Effects
    2. Negative Effects
    3. Neutral or Unsure Effects

Positive Effects

Five reasons for positive effects:

1. Doing dangerous work.

2. Performing tedious work with high efficiency and accuracy.

3. Handling aspects of computing work.

4. Undertaking jobs human labor does not want to do.

5. Addressing labor shortages and reducing business costs.

Negative Effects

Two reasons for negative effects:

1. AI and robotics taking over human labor, creating unemployment crises.

2. Replacing both manually repetitive and cognitive jobs, causing job insecurity.

Neutral or Unsure Effects

Various unique arguments for neutral or unsure effects:

- AI and robotics can both substitute and complement human labor.
- Total replacement of human labor is not yet feasible.
- Consumers prefer human-provided products and services, especially with high symbolic value.
- Importance of human labor learning to collaborate with AI and robotics.

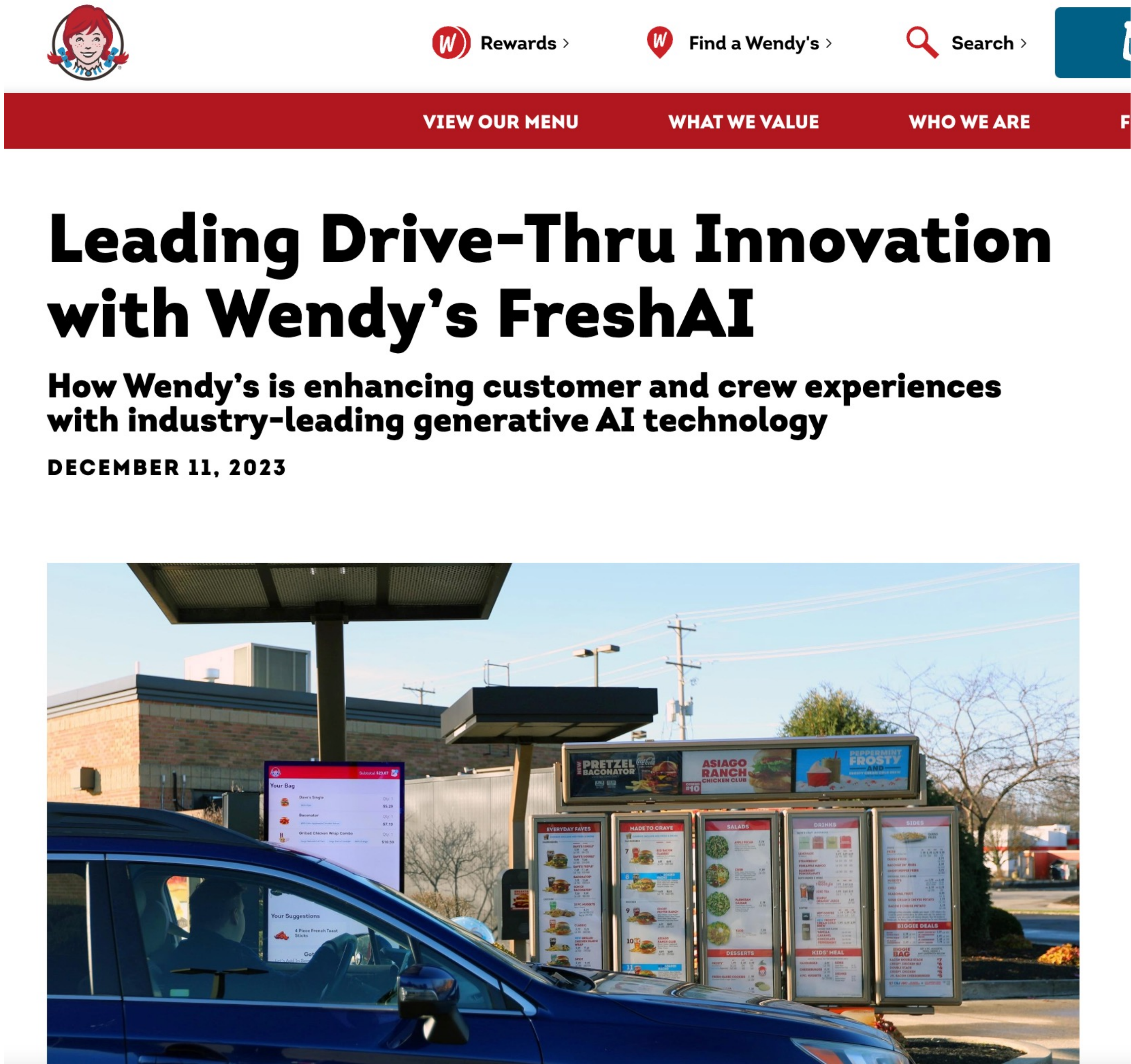
Conclusion

- AI and robotics impact various job sectors differently.
- Important to consider equity and fairness.
- Governments and businesses need to address potential job displacement.
- Future collaboration between human labor and AI is essential.

Impact on Entry-Level Jobs

- Impact on entry-level jobs, like AI in drive-thru ordering commonly held by racialized workers in Canada and the US. (April 2024)

https://www.wendys.com/blog/drive-thru-innovation-wendys-freshai



Leading Drive-Thru Innovation with Wendy's FreshAI

How Wendy's is enhancing customer and crew experiences with industry-leading generative AI technology

DECEMBER 11, 2023

4.3 My Research: Responsible AI: Perspectives from Canada and Norway

Research Project Overview

- Aim: Build groundwork to support global researchers and policymakers in instituting international agreements to regulate AI technologies.
- Successfully received research ethics approval from the Office of Research Ethics and Integrity, University of Ottawa.
- Current Status: Data collection phase for Canada and Norway. Plan to conduct in more countries, particularly low- and middle-income countries to obtain equity voice.

Importance of Diverse Perspectives

- Gathering perspectives from Norwegian and Canadian stakeholders.
- Involved stakeholders: Technologists and policymakers.
- Focus on the entire AI technology creation lifecycle.

Case Studies from Norway and Canada

- Results to be presented as case studies from Norway and Canada.

- Insights and lessons aimed to support global responsible AI principles.

- Main areas of focus: bias and fairness (discrimination), privacy, explainability, and trustworthiness.

5. Policy Recommendations

Advocacy

- Advocate for local and national governments to engage in the development of **international guidelines that establish best practices**, ensuring AI remains fair, unbiased, and non-discriminatory, and is deployed responsibly.

- Similarly to the example of Microsoft bans US police from using their AI facial recognition tool.

Collaboration

- Promote collaboration among tech companies, local and national governments, civil society organizations, and vulnerable populations to embed fairness, unbiased, and non-discriminatory in AI as **“a core societal and industrial framework”.**

شكراً (Shukran)

谢谢 (Xièxiè)

Thank you

Merci

Спасибо (Spasibo)

Gracias

Questions?

Channarong Intahchomphoo, Ph.D.
cintahch@uottawa.ca